Integrated Modeling of Complex Systems with Applications to Public Health and Sustainability

by

Gary Lin

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

October, 2018

Abstract

Understanding the dynamics of a changing world are of great interest to policy-makers, nonprofit organizations, governments, and businesses since society largely operates as a system. We develop system models to capture the complexity of the world in a logical and quantitative manner. Specifically, we use methods such as network analysis, time series analysis, system dynamics, and Markov Chains to explore systemic issues. These methods are applied to a socio-technical system related to public health and sustainability. We will also explore ways to capture this complexity by first identifying and analyzing the system with an interdisciplinary perspective then propose a method to integrate system models.

We begin by identifying the complexity of large-scale systems, such as Research & Development (R&D) of pharmaceutical treatments. In this project, we utilize a network representation to investigate collaboration among pharmaceutical companies and other stakeholders to determine the causes that enable success in developing a regulatory-approved therapeutic treatment. Secondly, we propose

ABSTRACT

an integrated multi-component model to capture the feedback loops that couples global population growth, environmental sustainability, and health systems. Finally, we investigate a system dynamics integration of a Markov Chain that describes migration patterns of the United States with respect to climate change.

Reader: Dr. Thomas Gernay

Reader: Dr. Sauleh Siddiqui

Advisor: Dr. Takeru Igusa

Acknowledgments

I am deeply grateful and privileged to work with an extraordinary advisor, Professor Takeru Igusa. Your guidance and generosity opened up many opportunities that allowed me to achieve more than I have ever imagined.

I want to thank Prof. Sauleh Siddiqui for his time on the thesis committee and reader – and thank you for supporting me, professionally and personally, through my academic journey. And, thank you Prof. Thomas Gernay for your openness and willingness to be part of my thesis committee and reader.

Thank you Prof. Stan Becker for investing so much time and effort in helping me succeed with my research.

Thank you Prof. David Rothman, Prof. Keith Porter, and Prof. John McCartney for their support during my undergraduate studies at the University of Colorado and helping me pursue my graduate career.

My doctoral research started in earnest with a project sponsored by the MIT Collaborative Initiatives. I learned about the importance of systems approaches and the roles of leadership from Dr. Tenley Albright, Director of the Collaborative,

which I appreciate deeply. I also want to thank others who worked with me on this project including Ellie Carlough and Pattie Lauria, also of the Collaborative, Jen Bernstein, Michele Palopoli, Viva Dadwal, Smile Indias, Felipe Feijoo, Francisco Del Canto, Patricia Natalie, and Lakisha Pooler.

Great thanks to my research and professional collaborators, Dr. Qingfeng Li, Prof. Lauren Gardner, Dr. Sundar Velkur, Prof. Rahmat Beheshti, Dr. Gonzalo Pita, Dr. Mehdi Jalalpour, Prof. Diego Martinez, Prof. Paul Locke, Justin Cooke, Prof. Danielle Wood, Prof. Seth Guikema, and Prof. Brooke Anderson.

I would like to thank Jimi Oke, Wei Jiang, Sriram Sankaranarayanan, Sen Lin, Zhaohao Fu, Qi Wang, Marietta Squire, Siyao Zhu, Chia-Hsiu (Todd) Chang, Fardad Haghpanah, Charalampos Avraam, and everyone in the civil engineering systems research group for the wonderful discussions and insights. Many thanks to Lisa Wetzelberger, Deborah Lantry, Amanda Jackson and Vess Vassileva-Clarke for their help throughout the years. I want to also thank my fellow colleagues, Anindya Bhaduri, Dave Fratamico, Hwanpyo Kim, Nicolas Venkovic, Aakash Bangalore Satish, Steven Chow, Tyler Tomita, George Weber, Farah Huq, Deniz Ozturk, Max Pinz, Xiaofan Zhang, Luigi Durand, Anna Scott, Yan Azdoud, Mattia Almansi, Elina Spyrou, CEGA, and the rest of the Civil Engineering Department. Finally, I would like to thank all my friends and family in my life – It truly takes a village to finish a Ph.D.

Financial support for projects and doctoral study was made possible by the

ACKNOWLEDGMENTS

MIT Collaborative Initiatives.

This material is also based upon work supported by the National Science Foundation under Grant No. 1331399. The support of the sponsor is gratefully acknowledged. Any opinions, findings, conclusions or recommendations presented in this thesis are those of the authors and do not necessarily reflect the view of the National Science Foundation.

This work received additional financial support from the Argosy Foundation, Bill and Melinda Gates Institute for Reproductive and Population Health, and Blomberg American Health Initiative Environmental Challenges Seed Grant.

Dedication

To Mom and Dad for all their love, support, and sacrifices.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"Everything must be made as simple as possible. But not simpler."

– Albert Einstein

In a world brimming with complexity, researchers must approach large-scale societal issues with a systems perspective. This includes socio-technical systems that consist of human beings and their behavior. A system containing humans produce chaotic and noisy signals due to the inherent heterogeneity in human decisions. Much like other natural phenomena, we utilize mathematical models to understand the main underlying mechanisms that govern these emergent behaviors. Models that are used to simulate complex systems in engineering can also be used to observe large-scale socio-technical systems that have many components and interactions. In this thesis, we will explore innovative ways of developing an integrated model that captures system-level complexity pertaining

to collaboration, sustainability, health, and population dynamics.

It does not take much imagination to characterize our world as a system with elements interacting with one another. However, systems can be viewed as "complex." The complexity of a system arise from the process of multiple components interacting with each other and its environment [1]. Components can be defined as individuals, populations, or physical systems that contribute to the emergent behavior of the system. The most famous example of complexity is the Lorenz attractor [2], which simulates atmospheric convection based on a deterministic system of three differential equations with three variables (x,y, and z). The variables are related to the rate of convection and temperature variation in two directions. The solution to the Lorenz attractor is chaotic and noisy which resembles a stochastic process, yet the system is evidently deterministic when you observe the phase space behavior. Figure 1.1 shows how such a simple system of three state variables cause complex behaviors that are commonly observed in the real world.

Similar signals are often observed in social systems where there are several sources of noise originating from interactions of humans. As a result, these signals should be isolated to determine how each component contributes to the noise and behavior of the entire system. Many traditional branches of science, engineering, and social sciences have observed these elements in isolation with great detail from deductive reasoning. As a systems modeler, it is their job to

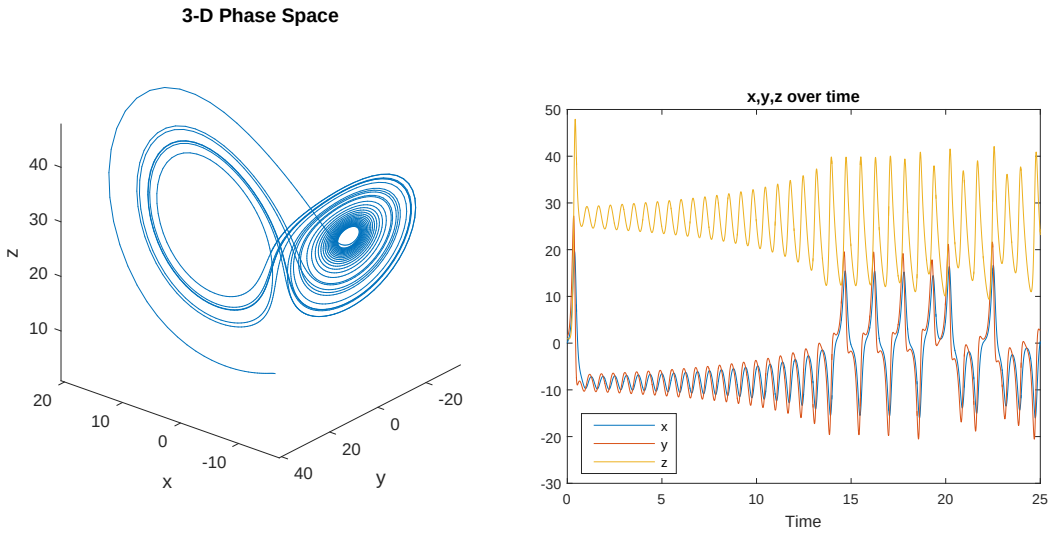


Figure 1.1: We recreated the Lorenz Attractor based on [2] to illustrate how simple systems may cause complex behaviors. The left figure shows the three-dimensional phase space of the system while the right shows a time series of the system for the three state variables x,y, and z.

deduce the basic mechanisms governing the emerging behavior and elicit the expertise of domain experts to develop models that encompass that knowledge. Furthermore, when one understands the deterministic mechanics underlying the measured signals, it is possible to infer from the model the causal pathways in social or health issues.

In the field of public health, the researcher is oftentimes faced with data that is complex and hard to decode the underlying phenomenon. They can use advanced machine learning or traditional statistical tools to identify the variables that are significant and impactful to the observed outcome. After delineating the factors that cause the behavior, they must seek out expert

opinion and other sources of information that potentially explain the behavior. Only after this process will the researcher effectively be able to develop an integrated systems model.

The steps needed to develop an integrated system model are outlined as the following.

1. identify the system by determining the *most* relevant components and interactions between the components,
2. develop a causal loop diagram that classifies components into causal links and measurable variables,
3. translate the causal loop diagram into a system model,
4. validate the model using empirical data, and
5. perform experiments, sensitivity analysis, and scenario analysis using the model.

It should be noted that the steps are not necessarily linearly and might require repeated iteration in order to develop a model that satisfactorily answer the research question.

With many new methods and tools such as machine learning and artificial intelligence, predictions have become more accurate and precise. However, these methods are still considered black-box methods that do not describe the

mechanistic interactions between different components of the system. We can utilize **system dynamics modeling** (SDM) to describe the causal relationships between elements of a complex system. System dynamics allow researchers to quantify and describe systems at multiple scales which range from industrial operations to world population growth. SDM has also allowed researchers across government, industry, and academia to observe emergent behaviors of complex system systems. System dynamics is considered to be "top-down" modeling approach to modeling a system.

With the increase of computational power, the ability to simulate systems from "bottom-up" also became feasible. The most popular method to simulating a system from the bottom-up is using **agent-based models** (ABM). This approach is still state-of-the-art and has been proposed as a straight-forward, and intuitive method to simulate the bottom-up mechanisms of human behavior. This method is strong at capturing the heterogeneity of agent behaviors. However, computing agent-based models may be expensive as well as tedious to parameterize given the nonlinearities and high-dimensionality. Another bottom-up approach uses a **Markov Chains** (MC) to simulate exclusive states of each agent based on a transitional probability between states. This is different from agent-based models because Markov chains are based on empirical observations whereas ABM is based on a decision function is based on theoretical assumptions with empirical calibration of the agent's decision function. Both bottom-up methods

have their advantages and drawbacks. ABM offers more information about the individual-level mechanism while MC offers computational advantages.

Both top-down and bottom-up approaches have real implications on the causal mechanism of socio-technical systems. In public health, interventions can also be characterized as bottom-up (e.g. grass-root campaigns, volunteerism) and top-down (e.g. regulatory policy, taxation). The ability to understand the system from these two scales is crucial to developing solutions for multifaceted public health issues since we are able to narrow down the mechanisms that contribute and exacerbate the systemic problems [3, 4]. Furthermore, complex issues are most likely not exclusively driven by one facet, but the interaction of multiple scales. For example, how do certain unhealthy behavior cause regulators to react? One such example of a multifaceted problem is childhood obesity [5–7] where systems thinking and modeling have contributed to understanding a complex, multiscale issue. Other public health applications that benefit from systems thinking include the environmental sustainability and health system operations [8–10]. All of these aforementioned problems include economic, demographic, social, and operational aspects that are usually studied in isolation by siloed disciplines that only study the system from a narrow perspective.

From a modeling standpoint, we can develop a framework that integrates both bottom-up and top-down models to represent high and low-scales of complex systems. Furthermore, we can also study one-way and two-way casual interactions

between components of different scales. Two-way interactions are also known as feedback loops, and more difficult to model than one-way interactions. An integrated model couples each component as a distinct "submodel." It should be mentioned that in order to capture the relevant complexity of the system, we need to be able to decompose the system into key features that are relevant to the scientific question at hand.

1.1 Translating system thinking into system models

Before we begin modeling, the researcher needs to distill the system into the relevant parts, we can isolate each subsystem as a separate component and analyze that particular component as a closed system with fixed inputs from other systems. By modularizing these components, each subsystem can then be modeled with different methodologies that are appropriate to the behavior of that subsystem.

The conception of an integrated systems model begins as a discussion among domain experts familiar with different aspects of the system in question. In order to pinpoint the causal structure of the model, collaboration from an interdisciplinary team of researchers with a priori knowledge about various aspects of the system is crucial. The discussion among an interdisciplinary group of researchers

should focus on observed and theoretical causal linkages that affect the systemic issue in question. One may develop a casual sketch similar to Figure 1.2. The development of these sketches does not need to be formal since many iterations are needed in order to synthesize a consensus understanding of the system that eventually leads to a quantitative model that is useful and descriptive.

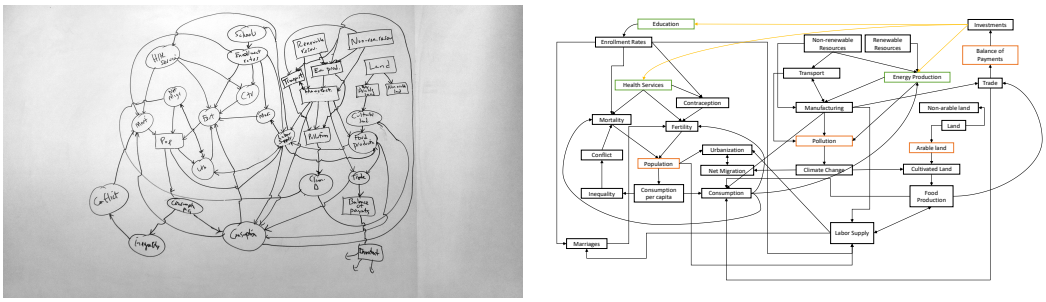


Figure 1.2: Examples of causal sketches on two different mediums. A systems approach requires an interdisciplinary discussion. Both diagrams attempts to capture system that impact population growth.

After the causal sketch is made, one can formalize this diagram as a **causal loop diagram** (CLD) [11]. There are other conceptual diagramming languages such as the **systems modeling language** (SysML) [12] that formalize systems thinking into a rigorous and logical framework. Furthermore, visualization also conveys a level of clarity that is easily accessible to a general audience. The causal loop diagram is developed based on linking variables that are causally related. This visualization method is depicted such that variables are considered nodes, and links are considered causal links. The links are unidirectional arrows that point from the influencer variable that (i.e., cause) to the influenced variable (i.e., effect).

Causal Loop Diagrams can be converted into a **system dynamics model**. System dynamics was developed by Jay Forrester [13–15]. System dynamics was originally meant to invoke systems thinking about a certain issue by endogenizing all relevant variables [16]. The system dynamics paradigm was originally created to be interdisciplinary in nature because it is conceptually intuitive. This allowed researchers without quantitative backgrounds to participate in the structural formation of the model.

Causal links in the a CLD are classified as positive and negative polarities. A positive causal link has a positive relationship. When the causal variable increases the value of the effect variable also increases. Negative polarity implies that when the cause variable increases, the effect variable *decreases*. For example,

$$\textbf{Positive Polarity}: \quad \text{Fertility Rates} \uparrow \quad \xrightarrow{+} \quad \text{Population size} \uparrow,$$

$$\textbf{Negative Polarity}: \quad \text{Mortality Rates} \uparrow \quad \xrightarrow{-} \quad \text{Population size} \downarrow .$$

Based on this logic, we can develop a causal loop diagram of a system. A classic example of a population system is the Lotka-Volterra model of competing populations, also known as the predator-prey model [17]. In Figure 1.3, we see the Lotka-Volterra model formalized into a CLD.

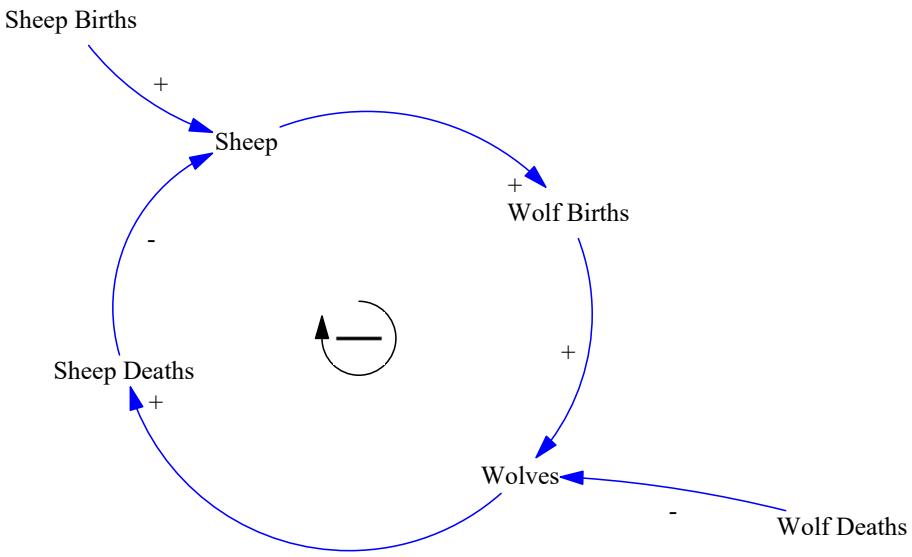


Figure 1.3: Lotka-Volterra model with showing the causal relationships between sheep (prey) and wolf (predator) populations. The polarity is indicated by the positive and negative next to the arrow heads. The loop in the center is an example of a balancing feedback loop, indicated by the negative clockwise arrow.

One of the purposes of causal loop diagrams is to identify feedback loops within the observed system. These feedback loops come in two variety: balancing and reinforcing. A balance feedback loop (also called a negative feedback loop) occurs when a set of variables counter each other. This happens when there is an inverse "cause and effect" relationship among a set of variables. Balancing loops tend to stabilize a system and keep the values from increasing or decreasing uncontrollably.

Reinforcing loops are the opposite of balance loops such that when one variable changes that influence another variable to change in the same polarity as the initial variable. This effect is then magnified in the system due to the circular cause and effect interaction. A good example would be money invested and the interest that it accrues – as the interest is reinvested into the principal amount, the interest that is earned in the future increases causing the principal amount to increase even more than before.

After understanding and diagramming the system, we can deduce the stocks and flows associated with the CLD by examining which variables accumulate (i.e., have a memory that is carried over each time step). If the variable is said to have memory (e.g., population size, a volume of liquid in a container), we will call this variable a **stock**. Stocks are adjusted by **flows**, which are variables that either increase (inflow) or decreases the stocks (outflow). Flows can be thought as a rate of change which is mathematically expressed as a

derivative. If we assume stock is y, and x^{IN} and x^{OUT} are respectively inflows and outflows, then we can mathematically express the stock as an ordinary differential equation (shown in Equations (1.1) and (1.2)). Variables that are not considered stock or flow, are called **auxiliary variables**.

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = x^{\mathrm{IN}}(t) - x^{\mathrm{OUT}}(t) \tag{1.1}$$

$$y(t) = y(t_0) + \int_{t_0}^{t} x^{\mathrm{IN}}(\tau) - x^{\mathrm{OUT}}(\tau)\ \mathrm{d}t \tag{1.2}$$

We convert the CLD of the Lotka-Volterra model in Figure 1.3 into Figure 1.4.

Once the model has been translated to a system dynamics framework where the stocks and flows, we will be able to build a mathematical model. The fidelity of the model relies on the ability to estimate the parameters that describe observational and experimental data. Different validation techniques are used to justify the fidelity of the model [18]. Also one should also be concerned about parameter estimation techniques since the parameter space tend to be highly nonconvex (one should refer to [19]). There are other nontrivial concerns like overfitting and optimizing the bias-variance tradeoff with dynamic models. However, parameter estimation and overfitting will not be the focus of this thesis. We will mostly focus on the development and integration of systems

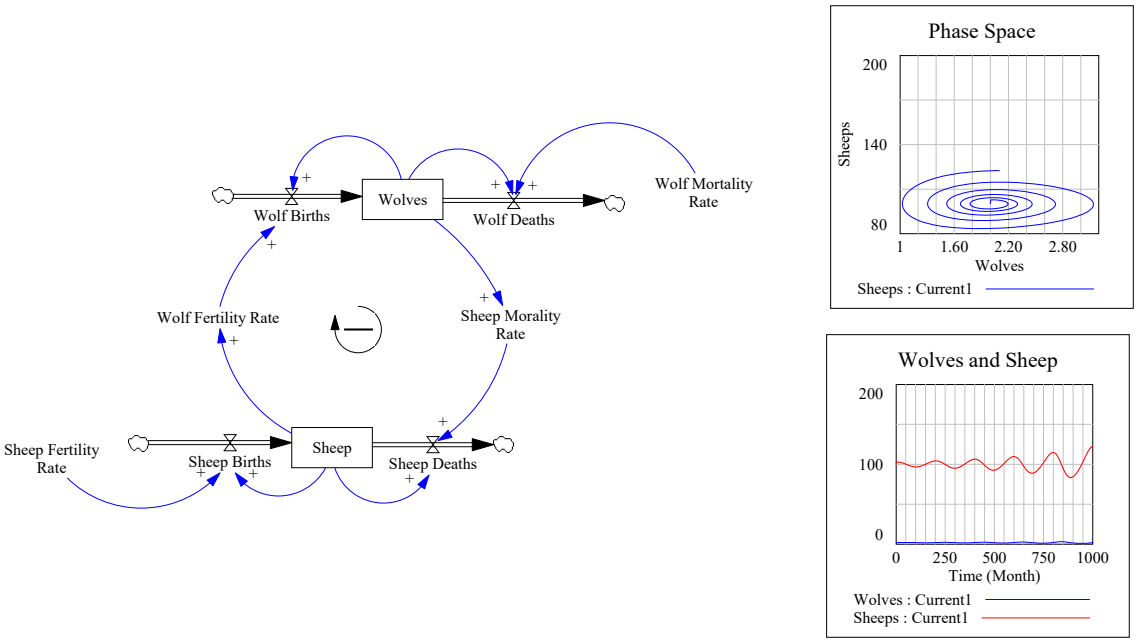


Figure 1.4: Lotka-Volterra model represented as a stock and flow. The stocks variables are *wolves* and *sheep*; inflows include *wolf births* and *sheep births* while the outflows include *wolf deaths* and *sheep deaths*. Auxiliary variables include *wolf fertility rate*, *wolf mortality rate*, *sheep fertility rate*, and *sheep mortality rate*. The behavior the model is also shown on the right side. This figure is produced using the Vensim software, which is one of the many commercial software that has an ordinary differential equation solver embedded.

models and tools.

1.2 Contributions and Applications of the Systems Approach

Many projects that are introduced in this thesis are interdisciplinary and ongoing at the time of writing. The goal of this thesis is to demonstrate how engineering methods and tools are used to develop integrated models of systemic issues. The field of engineering, especially civil engineering, has always encompassed an interdisciplinary approach to tackling problems. In addition, integrated systems modeling has been used widely in the field of civil engineering to analyze the coupled effects of infrastructure systems at multiple scales. This ranges from building systems (e.g. structural, heating, ventilation, and plumbing systems) to urban systems (e.g. transportation, drainage and sewage, water, and electrical grid systems). We can translate the same systems modeling tools into social and public health systems.

In Chapter 2, we present a collaboration network analysis of clinical trials, and the impact of network structure, actor, and partnerships has on the successful development of pharmaceutical therapeutics. The work spearheaded by the Johns Hopkins University Center of Systems Science and Engineering (JHU-CSSE) and the Massachusetts Institute of Technology Collaborative Initiative (MIT-CI) to investigate inefficiencies and issues with the process of clinical trials by including a team of interdisciplinary researchers. In addition to the

quantitative analysis that we presented, field research efforts such as interviews of stakeholders, round-tables, and ethnographic field visits were carried out in order for the team to determine the major barriers that hindered the clinical trials process. Although the regression model used in our network analysis is not integrated, it must be mentioned that this was the research design was the synthesis of an intensive deep dive into the complex system of pharmaceutical clinical trials. The identification of relevant research problems are a *necessity* for developing integrated models. The scientific contribution includes using an interdisciplinary approach that involves actual stakeholder discussions and field visits to study the clinical trials process. Other studies have investigated the collaboration network of pharmaceutical companies; however, we captured multiple types of actors in different stakeholder groups in our network analysis that ranges from non-profit organizations to academic institutions. We also make contributions to our understanding of the complexity in the clinical trials process by observing collaboration on a clinical level.

In Chapter 3 includes work sponsored by the Bill & Melinda Gates Institute for Population and Reproductive Health that was developed based on a research effort to understand the global system from a sustainability perspective, as well as its effect on population dynamics in various income-regions of the world. Furthermore, we investigated the growing socioeconomic inequality gap within the populations. This question led to the development of an ambitious integrated,

multi-component model that embodied elements of economics, demography, climate science, food and nutrition, education, health, and natural resource studies. These elements are classified into seven submodels in which the structure was developed based on consultation from various domain experts. We formulated the casual loop diagrams as well as the equations of motions based on well-established theories from social sciences and expert consultations. We also locate the data that would inform the parameter estimation of the model and propose a heuristic method of calibration of various submodels. In the final section of the chapter, we will introduce a road-map for a complete integration of the overall model. The contribution of this model to other integrated assessment models (IAM) includes an endogenized population submodel with fertility and mortality rates being affected by other submodels. Furthermore, we attempt to capture the inequity and disparities of resources, food, and health access of two socioeconomic groups, “rich” and “poor,” as well as between different income regions.

Finally, Chapter 4, we introduce a Markov Chain Model embedded into a system dynamics model that contains a coupled feedback between population dynamics such as migration, and climate change. This work was based on a research effort to understand urbanization, public health, and climate change. This study was funded by the Bloomberg American Health Initiatives (BAHI) Environmental Challenges Seed Grant to explore public health issues in the

United States related to the environment. The scientific contribution of this project includes a Markov Chain integration embedded in a system dynamics model.

Integration can vary case-by-case depending on the application. In Appendix A, how one can couple two models of different scales. This project is affiliated with an NSF Hazard-SEES funded project that originated from an interdisciplinary research effort to develop an integrated model of repeated hazards and infrastructure resilience. The integrated model utilizes an agent-based model (ABM) and climate model, and our intermediate model provides an interface for the two models. We developed the intermediate model using time series analysis and is mathematically formulated as a distributed-lag regression that predicts indoor temperature based on outdoor temperature. We demonstrate the utility of the model against actual data collected from a single house in Baltimore, Maryland, USA.

1.3 Thesis Summary

In this thesis, we will focus on model identification of complex systems and the integration of system models to capture complex behaviors. All of these models were formulated based on an interdisciplinary effort to understand large-scale, complex issues that overlap into multiple scientific domains. We

specifically look at the research of medical R&D networks, heatwave, and population and sustainability, and climate-driven migration – all systemic issues that require the researcher to employ the systems approach. We demonstrate the system method in the following.

Chapter 2 : Distilling Complexity of Collaboration Networks in Clinical Trials – Identify complexity of a system and relevant components

Chapter 3 : Multi-component Integration of System Dynamics with Applications to World Population Growth and Sustainability – Propose a conceptual framework and formulate a multi-component, integrated model

Chapter 4 : Integrated Markovian Modeling of Climate-driven Migration and Urbanization in the United States – Integrate a model with a feedback loop

Chapter 5 : Conclusion and Outlook

Chapter 2

Distilling Complexity of Collaboration Networks in Clinical Trials

We investigate the complexity of clinical trials system by identifying a network of collaborations and the impacts of that network on the success of developing a drug therapy that is approved by the US Food and Drug Agency (FDA). The purpose of this chapter is to demonstrate a network characterization of a complex problem in the medical research field. In the past, business and public health researchers have looked at the system in parts. However, this system was studied extensively as a part of an interdisciplinary team to determine barriers and issues that impact an inefficient system.

2.1 Background

Drug research and development (R&D) is an incredibly complex, expensive undertaking which is prone to failure. Given that on average it may take over 10 years and cost up to 2.6 billion dollars to develop a single approved molecule [20], drug R&D has become a collective effort. During a drug's lifespan, it is common for a spectrum of actors from government, academic, and nonprofit organizations to pharmaceutical and biotechnology companies to conduct phases of the basic, preclinical, and clinical research – each contributing towards the development of a drug that is eventually approved by a regulatory agency. Furthermore, these actors collaborate to increase their research capabilities through access to key technologies or specialized knowledge developed and/or possessed by other actors [21–23]. They also often form vertical alliance networks where each actor performs a relatively distinct set of activities along the value chain [24]. More than ever, success or achievement in drug development is dependent upon these collaborative networks; therefore, it is useful to examine the collective structure of these collaborative networks and the actors involved in order to understand the impact of collaborations on the drug R&D process.

One can argue that the system is considered a complex, thus requires researchers to be able to quantify the complexity and model the attributes of the system that contribute to success. Network analysis offers a particularly useful set of

tools for examining both the structure of collaborative networks which exist to develop drugs as well as the combination of actors who bring them to market. In this body of literature, researchers examine network "nodes", i.e., the for-profit pharmaceutical and biotech companies as well as non-profit academic and government actors, and the collaborations between them called "links." Some studies have examined collaborative networks based on contractual alliances within the pharmaceutical industry [25,26], while others have studied knowledge networks which map the dissemination of knowledge via patent citations [27–30]. Other analyses have considered the interdependence between organizations' geographical locations [31] and their network position. The goal of these analyses is to measure the impact of collaborative networks on both the creation and acquisition of knowledge as well as actual drug development and approval.

Previous research has determined that "cohesion," a basic network property used to characterize the structure of networks, impacts the speed and reach of knowledge transfer between actors which in turn affects productivity. Cohesion may be thought of as the level of connectivity in a network. Cohesion is measured by the *average path length* and *global clustering coefficient* [29, 32]. Studies suggest that a low cohesion network, a network with few connections, that has a low global clustering coefficient results in a larger path distance between each actor (see Fig. 2.1). In a weakly connected network with low cohesion, actors would not be able to transfer information efficiently since information

would likely have to travel through more actors to get to another actor. On the other hand, a highly cohesive network may reach a point where excess connections lead to frequent transfers of redundant knowledge which in turn hinders research productivity [33–35]. This theory is referred to as the "echo chamber effect" and is supported by previous studies [36,37]. The implication is that when actors who have similar experience transfer knowledge repetitively to the same set of actors, they may reduce the marginal production of knowledge and inhibit research productivity.

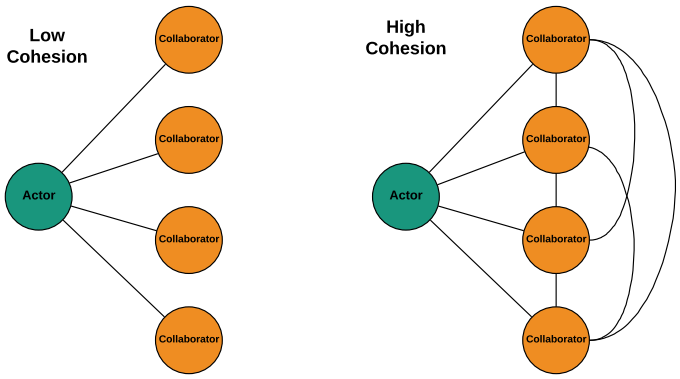


Figure 2.1: The local network of two actors are shown. Both actors are each connected to four collaborators (other actors in the network). The actor on the left has a lower cohesion with a local clustering coefficient of 0 which is the minimum possible value. The actor on the right has a higher local clustering coefficient of 1 which is the maximum possible value.

In addition to cohesion, existing studies suggest the organizational characteristics of the actors and their collaborators within the network, such as type of actor and portfolio (i.e., range of approved and candidate drugs held by an actor), also play an important role in knowledge acquisition and creation, which contributes

to higher clinical research output (e.g., the number of drugs approved, research efficiency, etc.). Research indicates there are advantages to partnering with a diverse network of collaborators [38–41]. One reason is that collaboration with a partner that has a uniquely different knowledge base and portfolio allows an actor to explore domains that were previously outside its own expertise and potentially difficult or impossible to access without a knowledgeable partner [22]. Another reason a diverse network may be useful is that a wider knowledge base might help actors maintain alliance ties [42]. Additionally, research indicates actors may seek to obtain or exploit innovations developed by partners [23], but to truly innovate, an actor must be able to combine pre-existing knowledge with new knowledge that was obtained through collaboration [28]. Therefore, diverse collaborations through networks comprised of heterogeneous actors may serve to expand an actor's knowledge base and/or portfolio which may be beneficial for drug research and development.

In the present study, we characterize the actors and networks involved in drug development and examine the extent to which both network cohesion and diverse collaborations foster the development of new drugs. Rather than consider the networks formed through patents or contractual alliances, this is the first study to examine the networks formed when actors collaborate to run clinical trials in order to bring drugs to market. Using data from both publically-available and proprietary sources, we followed the shifting collaborations

of 4,494 actors who sponsored a trial or partnered in the execution of a trial over a ten-year time period. We found that actors positioned in networks that have low cohesion tend to fare better in developing drugs that eventually receive regulatory approval. In other words, actors that collaborated with partners that were not connected with each other tend to succeed more often than actors that were more embedded in dense clusters of collaborations. Furthermore, actors that were less embedded in the network tend to collaborate with a wider range of partners that have varying expertise in different therapeutic areas. Therefore, actors are more likely to participate in conducting a clinical trial for a drug that is later approved when they collaborate with a diverse set of partners.

2.2 Methodology

2.2.1 Constructing a Clinical Trials Collaboration Network

We focused our investigation on empirical data from 4,494 organizations and 18,040 trials extracted from Aggregate Analysis of ClinicalTrial.gov (AACT) and BioMedTracker Pharma Intelligence databases (See S1 for data collection and processing methods). Using this data, we were able to construct a two-

mode affiliation network that included the actors (sponsors and partners) and clinical trials (distinguished by the national clinical trial (NCT) identifier number). We then transformed the two-mode affiliation network using a bipartite projection into a one-mode collaboration network that ranged from January 2006 to January 2016. On a one-mode network, a node represents a single actor, and a link represents at least one instance of collaboration on a clinical trial between a pair of actors. We conducted an egocentric analysis on a dynamic, one-mode collaboration network where the focus is on the organizations or *actors* over a period of time. The network representation allows us to observe patterns of collaborations that facilitate the transfer of knowledge (See S2 of SI).

In order to account for temporal differences, we generated multiple time-dependent networks to capture the evolution of the network structure over the considered time period. The network was generated for each month between January 2006 to January 2016, which resulted in 121 snapshots of the clinical trials system. A node or link is active at a given snapshot if the respective actor and collaboration are involved in at least one trial or collaboration during that month. Based on this network, we developed several metrics to measure organizational, collaboration, and structural characteristics of actors in the clinical trials system for each month. We then performed a lagged regression analysis on our calculated metrics to relate success.

2.2.2 Network Measures and Actors' Attributes

We examined the success of each company in terms of research output and productivity. The metrics that we used are *cumulative trial successes* (CTS_{it}) and *trial success rate* (SR_{it}). Cumulative trials success is defined as the cumulative number of clinical trials that an actor has been involved in as a sponsor or collaborator that eventually led to an FDA-approved drug. The trial success rate is simply the cumulative trial successes divided by the overall number of trials that an actor has been involved in as a sponsor or collaborator. The trial success rate captures the organization's effectiveness in achieving its research objectives.

We classified the actors into six organization types: *academic*, *government*, *nonprofit*, *industry*, *hospital system*, and *large pharmaceutical companies*. We determined the classifications based on additional data gathering efforts using publicly-available sources and other methods (See S3.1 in SI). This allowed us to stratify and add in fixed effects to control for the organization type in our regression. We wanted to distinguish the organizational type because each actor plays a distinct collaborative role in a clinical trial

For each actor on the network, we computed several node-specific metrics that quantified expertise, structural, organizational, and collaboration characteristics. Expertise is determined by designating each actor as having a specialization in one particular therapeutic area (See S3.2 in SI). Structural characteristics are

characteristics of the local network of the actor (See S3.3 in SI). Organizational characteristics are based on the clinical research experience of the actor (See S3.4 in SI). Collaboration characteristics are based on the relative comparison of all collaborators versus the observed actor (See S3.5 in SI). In the following of this section, we will only present selected variables that were used in our regression. Please refer to the Supplementary Information for a complete listing and description of all measures.

Tomasello et al. [43] defined the knowledge distance as the Euclidean distance between organizations i and j at time t. In other economic literature, this is known as the technological distance [44]. This is formally defined as

$$KD_{ijt} = \|\mathbf{x}_{jt} - \mathbf{x}_{jt}\| = \sqrt{\sum_{d \in \mathscr{D}} (x_{idt} - x_{jdt})}. \tag{2.1}$$

where x_{idt} represents an element of the knowledge mix vector $\mathbf{x}_{it}$ (see Eq. 1 in S3.4 in SI) with element x_{idt} representing the fraction of clinical trials conducted in therapeutic area d at time t.

This link-specific metric was meant to compare the differences between the two firms' patent portfolios. However, we have adopted this metric to measure the research differences between a pair of organizations' portfolios (i.e. the distribution of trial experience in each therapeutic area). The knowledge distance is at a maximum ($KD_{ijt} = \sqrt{2}$) when actors are concentrated in two exclusively, different therapeutic areas. When two firms are concentrated in the same

therapeutic area, the knowledge distance equals to 0 because they are identical in expertise while a higher KD corresponds with a larger difference.

One of the properties of the Euclidean-based knowledge distance in (B.3) is that the measure takes into account research diversification of an actor. Let's say Actor 1's portfolio is solely concentrated in Neurology, Actor 2 is solely concentrated in Urology, and Actor 3 is divided between Oncology and Urology. In this situation, the knowledge distance between Actor 1 and Actor 2 is larger than Actor 1 and Actor 3 even despite the fact that both Actor 2 and 3 are in exclusive research fields relative to Actor 1. This is a well-known property of Euclidean Distances and fits our case since we are implying that actors that are more interdisciplinary have more capacity to function in other fields than specialists.

In our analysis we calculate the *mean knowledge distance* $\langle KD \rangle_{it}$ for all incident links to actor i at time t and used it as a variable in our regression. We can define this as

$$\langle KD \rangle_{it} = \frac{\sum_{\{i,j\} \in E(G)} KD_{ijt}}{\delta_{it}} \quad s.t.\ i \neq j$$

where δ_{it} is the number of degrees for actor i at time t.

An actor may decide to adopt two approaches: broadly diversify in different therapeutic areas (i.e., "jack-of-all-trades") or specialize in one therapeutic area (i.e., "master-of-one"). The decision to diversify is usually driven by the size

of the organizations which determines the economies of scope [45]. Larger companies tend to diversify more than smaller companies. Furthermore, the firm must decide whether to collaborate with a jack-of-all-trades actor or a master-of-one actor that has a deeper specialization in one field.

We quantified *research diversification* using an entropic measure that measures the heterogeneity of actor i's knowledge mix $\mathbf{x}_i$.

$$RD_{it} = \sum_{d \in \mathscr{D}} x_{idt} \ln \left(\frac{1}{x_{idt}} \right) \tag{2.2}$$

where x_{id} is the element of the knowledge mix vector that represents the percentage of clinical trials experience in disease d. This measure gives us an impression of the level of interdisciplinary in an organization's portfolio. We assume that a company that has completed 0 trials will have an entropy of 0.

From the research diversification that is defined in Equation (B.2), we can determine *mean research diversification* of all the neighboring organization of actor i at any given time period t. The mean research diversification $\langle RD \rangle_{it}$ is simply

$$\langle RD \rangle_{it} = \frac{\sum_{\{i,j\} \in E(G)} RD_{ijt}}{\delta_{it}}.$$

We were able to capture local cohesion or embeddedness of each actor using

the *local clustering coefficient* which is commonly defined as

$$CC_{it} = \frac{2L_{it}}{\delta_{it}(\delta_{it} - 1)}, \tag{2.3}$$

where L_{it} represents the number of links between the neighbors of actor i and δ_i represents the degree (number of partners) of actor i at time t. The local clustering coefficient is a good indicator for cohesion since it can show the extent to which an actor is embedded in the treatment development system. Local clustering coefficient specifically depicts the connectivity within an actor's set of partners.

In our study, we have assumed knowledge is developed through an actor's experience which can be quantified as the number of trials conducted in each therapeutic area (e.g. Neurology). By stratifying knowledge by therapeutic area, we can determine the relative competencies of each actor in the network and measure the extent to which their knowledge is concentrated or distributed. Furthermore, we can designate each actor as an "expert" in a particular field by observing where most of their trial experience has occurred.

Once we designated each actor as an expert in a particular therapeutic area, we measured *collaboration diversity* by using an entropic measure of diversity,

$$CD_{it} = \sum_{d \in \mathscr{D}} z_{idt} \ln \left(\frac{1}{z_{idt}} \right). \tag{2.4}$$

The variable z_{idt} is the number experts in therapeutic area $d \in \mathscr{D}$ that actor i is actively collaborating with at time T. The set $\mathscr{D}$ includes all the therapeutic areas that were defined in the BioMedTracker Pharma Intelligence database. This entropic measure is commonly used to quantify diversity in many fields that range from biology to production portfolios [46, 47]. This metric examines the set of partners for each actor and measures the diversity of expertise. In essence, collaboration diversity characterizes the breadth of expertise in the actor's collaboration network. Actor A's collaboration diversity would be ~ 0.64 if they worked with 2 experts in neurology and 1 expert in oncology. Actor B would have a collaboration diversity of ~ 3 if they conducted 20 trials with 20 different experts all specializing in different therapeutic areas. In this case, we would say Actor B is more diverse in their collaboration.

2.2.3 Multivariate Regression Analysis

We conducted a regression analysis on two response variables that relate to research output and productivity: cumulative trial successes and trial success rate. We ran separate regressions on each response variables with 1, 2, and 5-year lag to capture the delay of knowledge adoption and implementation. In order to demonstrate the robustness of our regression analysis, we conducted three separate sets of regression: (*i*) regression with only the control variables, and (*ii*) regression with selected variables, (*iii*) regression with all measured

variables (See S4 in SI). The regressions with selected variables are of interest since they are the most parsimonious models with minimized multicollinearity and statistically significant variables.

The lagged regression that examined cumulative trial successes (CTS_{it}) with respect to each actor i at time period t utilizes a negative binomial generalized linear model. We chose a negative binomial generalized linear model because the distribution of CTS_{it} was overdispersed. The regression with selected variables is defined as

$$\begin{aligned}\log(CTS_{it}) = \beta_0 &+ \beta_1 PrevSucc_{i(t-k)} + \beta_2 Trials_{i(t-k)} \\ &+ \beta_3 CD_{i(t-k)} + \beta_4 \langle KD \rangle_{i(t-k)} + \beta_5 CC_{i(t-k)} \\ &+ \gamma_t + \kappa_i + \epsilon_{it}. \end{aligned} \tag{2.5}$$

The explanatory variables include cumulative trials conducted $Trials_{i(t-k)}$, collaboration diversity $CD_{i(t-k)}$, local clustering coefficient $CC_{i(t-k)}$, and mean knowledge distance $\langle KD \rangle_{i(t-k)}$. We also control the model with a dummy variable, $PrevSucc_{i(t-k)}$, which takes on a value of 1 if the actor has achieved at least one success before time $t-k$. The fixed effects for time γ_t and actor-type κ_i are also included in our analysis. The response variable CTS_{it} is lagged k years, therefore all the covariates corresponding with each actor are at an earlier time $t-k$.

Since the trial success rate SR ranges from 0 to 1, we used a lagged beta regression. Beta regressions are only used to predict values in the (0,1) range, therefore a transformation was conducted on SR (see section S4.2 in SI) to convert the 0 and 1 values to be within the prescribed range. The trial success rate is defined as the cumulative number of trial successes normalized against cumulative number of trials. The regression can be shown as

$$\begin{aligned} g(SR_{it}) = \beta_0 + \beta_1 PrevExp_{i(t-k)} + \beta_2 \langle KD \rangle_{i(t-k)} \\ + \beta_3 CC_{i(t-k)} + \beta_4 RD_{i(t-k)} + \beta_5 \langle RD \rangle_{i(t-k)} \\ + \gamma_t + \kappa_i + \epsilon_{it}. \end{aligned} \tag{2.6}$$

The dummy variable $PrevExp_{i(t-k)}$ takes on binary values and represents whether the actor has conducted at least 6 trials before time $t-k$. The variable $\langle KD \rangle_{i(t-k)}$ measure the mean knowledge distance between the actor and collaborators. The knowledge distance is a Euclidean distance between two actors' research mix vectors. This measure was meant to capture the similarity in knowledge bases between a pair of actors. Collaboration diversity was not included in (B.6) because it was not statistically significant for predicting success rate. $RD_{i(t-k)}$ and $\langle RD \rangle_{i(t-k)}$ quantify research diversification and mean neighbor's research diversification, respectively. Similar to collaboration diversity, the research diversification is an entropic measure that quantifies the breadth of an actor's

research activity in each therapeutic areas.

The variables $Trials_{i(t-k)}$, $PrevSucc_{i(t-k)}$, $PrevExp_{i(t-k)}$ and the fixed effects γ_t and κ_i in (2.5) and (B.6) are considered to be the control variables. Robust checks include conducting regressions on control and all measured variables for all three time lags. The results of these other models are located in the supplementary tables in Appendix B.

2.3 Results

The network characteristics of different actors varied among organizational types. Academic actors tend to have more collaborative links across all organization types (See Fig. 2.2). This is not surprising since academic centers offer resources and infrastructure for clinical trials that are not available to other actor-types. As a result, academic collaborations tend to dominate this collaboration network in terms of connections. Many principal investigators on clinical trials also have an academic appointment, even if the trial is sponsored by industry, which explains the higher count of academic actors.

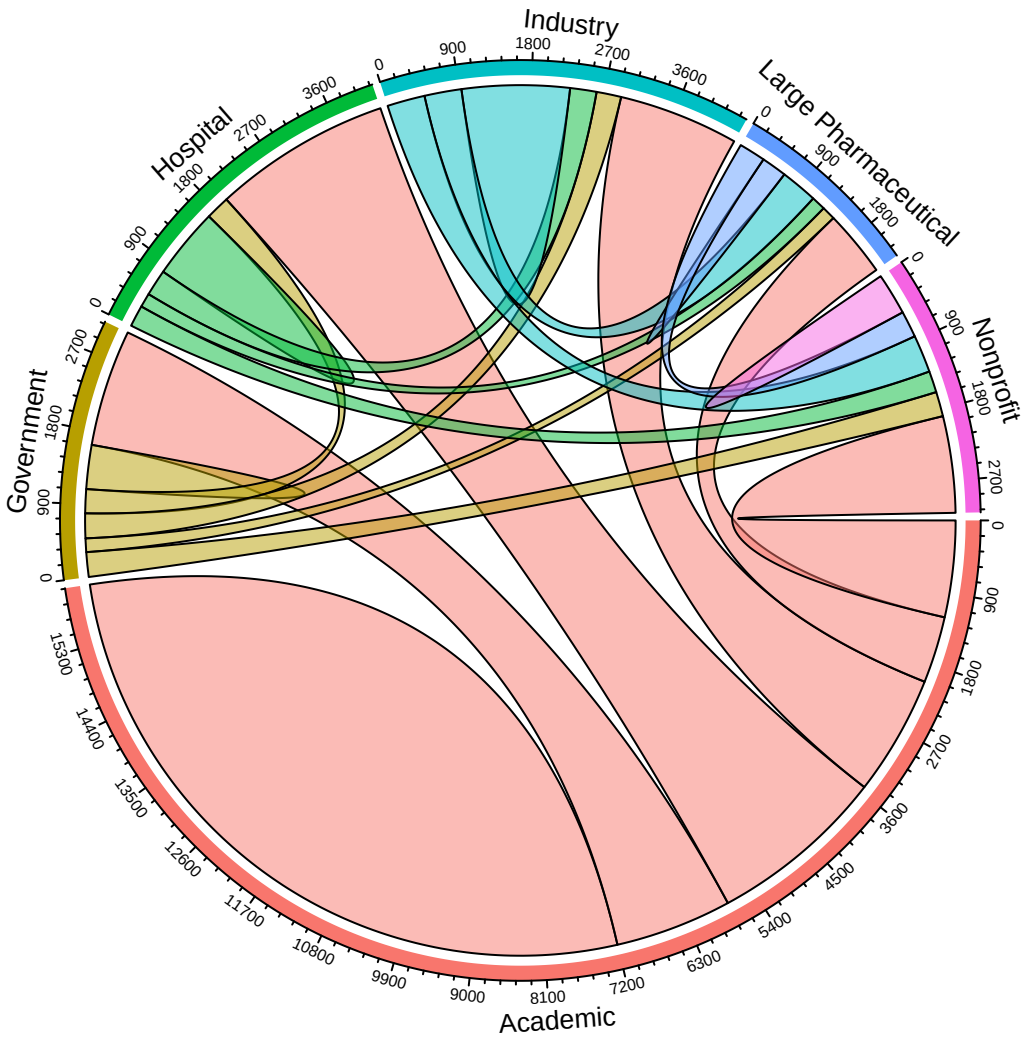


Figure 2.2: This chord diagram illustrates the number of collaboration links between and within the six organizational-types in January 2015. The width of the links is scaled based on the volume of collaborations.

	Cumulative trial successes			Trial success rate		
Variable	1-year lag	2-year lag	5-year lag	1-year lag	2-year lag	5-year lag
Previous success	2.709^{***}	2.177^{***}	1.285^{***}			
	(0.036)	(0.031)	(0.046)			
Previous experience				0.236^{***}	0.196^{***}	0.077
				(0.052)	(0.061)	(0.105)
Mean knowledge distance	-0.330^{***}	-0.263^{***}	0.045^{**}	0.025^{*}	0.034^{**}	0.031^{*}
	(0.016)	(0.017)	(0.020)	(0.014)	(0.014)	(0.018)
Cumulative trials conducted	0.331^{***}	0.301^{***}	0.236^{***}			
	(0.007)	(0.008)	(0.015)			
Collaboration diversity	0.240^{***}	0.270^{***}	0.269^{***}			
	(0.015)	(0.015)	(0.021)			
Local clustering coefficient	-0.075^{***}	-0.084^{***}	-0.154^{***}	-0.037^{***}	-0.041^{***}	-0.040^{**}
	(0.014)	(0.014)	(0.019)	(0.014)	(0.015)	(0.018)
Research diversification				0.092^{***}	0.103^{***}	0.102^{***}
				(0.019)	(0.020)	(0.026)
Mean neighbor research diversification				0.100^{***}	0.115^{***}	0.110^{***}
				(0.014)	(0.014)	(0.017)

Table 2.1: The standardized coefficient estimates with standard errors are shown for response variables, cumulative trial successes and trial success rate. The coefficients for cumulative trial successes is estimated by the negative binomial regression, while the coefficient estimates of trial success rate are the result of a beta regression. The statistical significance are indicated as $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ and standard errors are in parenthesis.

The estimated coefficients for the regressions with selected variables against cumulative trial successes and trial success rate are listed in Table B.11. For regressions estimating cumulative trial success, we found that the local clustering coefficient is a significant indicator with $P < 0.01$ in most cases and negatively correlated with research output. This indicates that as an actor 's network becomes more cohesive, cumulative trial successes and trial success rate actually decline. For all the regressions, local clustering coefficient has a significance of at least $P < 0.05$ (See SI for a complete listing of effect sizes and P-values).

Collaboration diversity is a significant explanatory variable for cumulative trial successes with $P < 0.01$ for all regressions. When we look at trials success rate, collaboration diversity was only significant for a 5-year lag with $P < 0.1$, thus not included in the regression with selected variables.

In our analysis, we noticed that collaboration diversity did not lead to higher research productivity. In the regressions against success rate SR with all measured variables, collaboration diversity exhibited a slight negative coefficient estimate and is not significant in most cases (see Models B1-3, B2-3, B3-3 in SI Tables 9-11).

Mean knowledge distance is significant for both cumulative trial successes and trial success rate, which we mentioned as an average measure of portfolio similarity between an actor and its collaborators. The regression coefficient value of mean knowledge distance for cumulative trials success is negative for

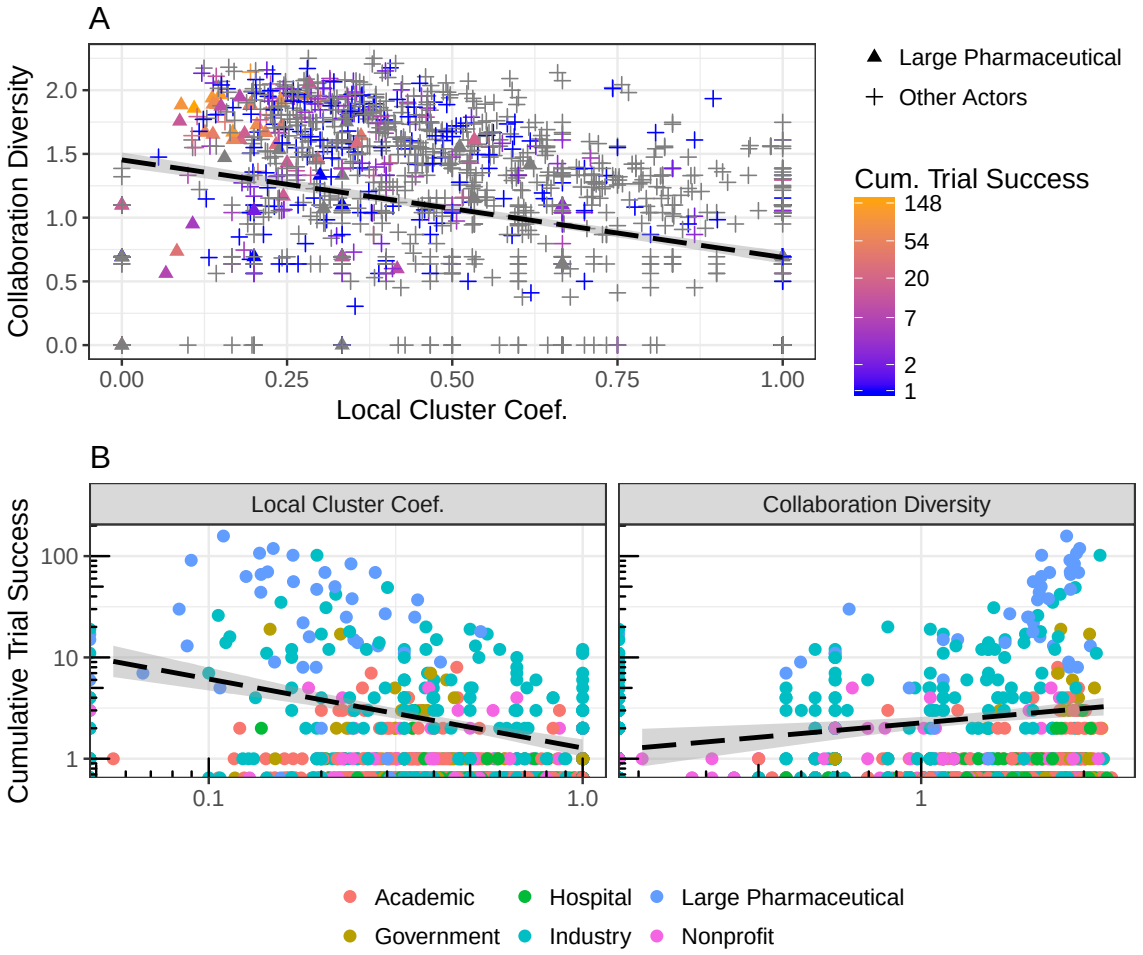


Figure 2.3: *(A)* Scatterplot of collaboration diversity versus local clustering coefficient in January 2015. Large pharmaceutical companies that have more success are distinguished as triangles. The black dashed trend line shows a linear negative correlation between the clustering coefficient and collaboration diversity. The color gradient represents cumulative trials successes. This plot highlights the research performance relative to collaboration diversity and local clustering coefficient. The gray points represent actors that are active for that period but have not subsequently participated in a successful clinical trial. *(B)* The scatterplot showing the relationships of cumulative trials success with respect to local clustering coefficient and collaboration for January 2015. Each organization-type is distinguished by color.

1 and 2-year lags which suggests that there is a negative correlation between research output and dyadic similarity, while the mean knowledge distance has a positive effect at the 5-year lag. In contrast, mean knowledge distance has a strictly positive impact on trial success rate. Considering collaboration diversity and mean knowledge distance, we can deduce that actors that collaborate with a set of partners that are not diverse (more concentrated in one field) but different from the focal actor's own expertise will be more likely to succeed.

We found that research diversification and mean neighbor's research diversification are significant factors toward the trials success rate with coefficient estimates being both positive. This suggests that organizations that have diversified their research activities tend to be more productive. Large pharmaceutical companies tend to have higher research diversification in our dataset than other actors.

Based on the regression results, we dove deeper into the relationship between cohesion and collaboration diversity. In Fig. 2.3A, we noticed there is an inverse relationship between collaboration diversity and local clustering coefficient which is indicated by the negative-sloped linear trend line. This indicates that actors that collaborate more diversely are less embedded in the network. In Fig. 3B, we examine this finding in more detail. Most actors with lower local clustering coefficient and higher collaboration diversity tend to be large pharmaceutical actors that have achieved more research output. Large pharmaceutical companies

collaborate with many smaller, new market entrants that are not as embedded in the collaboration network, like biotechnology start-ups [48]. As a result, large pharmaceutical companies are proficient at absorbing distinct knowledgebases of smaller, more specialized actors resulting in a lower clustering coefficient.

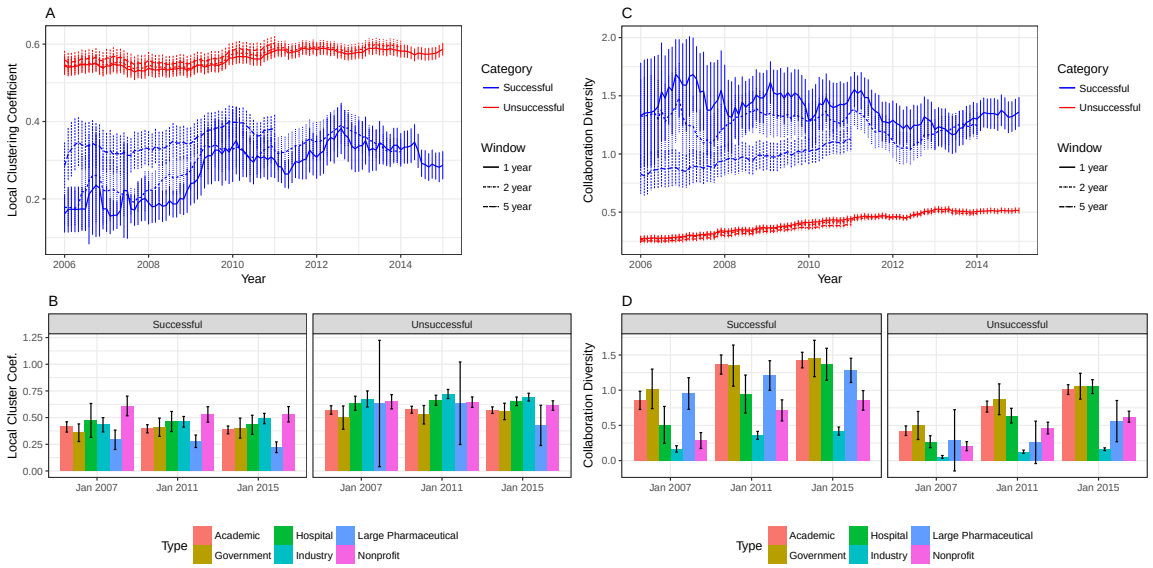


Figure 2.4: *(A)* The plot shows the average local clustering coefficient for dynamically defined successful and unsuccessful actors in the network at each month. Successful actors are dynamically defined at each time period t as organizations that will achieve at least one successful trial within the forward time window ranging from t to $t + k$, where k is the lag length. Otherwise, the organizations are characterized as unsuccessful. The standard deviations are shown as error bars. *(B)* In the figure, we divide into two static sets: *successful actors* ($n = 888$) and *unsuccessful actors* ($n = 3,606$). The average local clustering is stratified by actor-type and success for 3 time periods. The error bars show the 95% confidence interval. *(C)* Average collaboration diversity of dynamically defined successful and unsuccessful actors. *(D)* Same as (B), the static set of unsuccessful and successful actors' average collaboration diversity among different actor-types.

Fig. 2.4A shows the differences in clustering coefficient stratified by successful and unsuccessful actors as well as organization-types. Successful actors tend

to have a lower clustering coefficient than their unsuccessful counterparts. If we focus on organization types in Fig. 2.4B, we will notice that nonprofit organizations tend to be more embedded in the system, which indicates their role as a collaboration maker in the system. However, even the successful nonprofit versus unsuccessful nonprofit organizations have observable differences in their level of cohesion.

Fig. 2.4C and Fig. 2.4D shows the average collaboration among different groups of actors that are stratified by success and actor-types. We notice government and academic actors are the ones with the most collaboration diversity. This is expected since academic and government institutions are usually responsible for leading and sponsoring many clinical trials across therapeutic disciplines. We also notice that successful large pharmaceutical companies are more likely to seek out a diverse set of actors. This supports existing evidence of biotechnology companies searching for novelty in knowledge from various scientific communities [49].

2.4 Discussion

In this analysis, the local clustering coefficient proves to be a significant variable that is negatively correlated with research output and productivity. In some instances, betweenness centrality is a significant positive indicator

of cumulative trial successes, which is generally inversely proportional to the local clustering coefficient (See Fig. 10 in SI for correlation plot). These two network metrics demonstrate the importance of network position for innovation. This finding concurs with the findings of [36], who found that local cohesion tends to hinder innovation since stakeholders are essentially stuck in an echo chamber.

We also find that collaboration diversity is related to a higher cumulative trial success rate and trial success rate. We suggest that there may be two different reasons why actors that participate in collaborations with diverse partners are more likely to develop a drug that leads to regulatory approval. The system shows signs of preferential attachment which occurs when actors are attracted to other actors that have a demonstrated record of competencies, which is sometimes known as the "rich-get-richer effect" [21]. Larger pharmaceutical companies that are historically successful benefit from preferential attachment such that they attract a wide range of actors with varying therapeutic expertise and experience. Fig. 2.3 shows this effect, where actors classified as large pharmaceutical companies tend to have lower clustering coefficients with more collaboration diversity.

Our findings around mean knowledge distance, demonstrates that actors must consider more than diversity when choosing collaborators. When two actors with lower knowledge distance collaborate, they improve research output

and productivity. This suggests that companies with similar portfolios benefit from working with one another.

We find that collaboration diversity is negatively correlated with the clustering coefficient, which indicates that cohesion may not facilitate diverse expertise. We hypothesize that actors benefit from exposure to diverse ideas and knowledge through clinical trials collaboration due to the knowledge exchanged with partners on the peripheries of the system. Large pharmaceutical companies oftentimes recognize the benefit of knowledge diversity and collaborate with diverse actors that complement their own competencies. This is reflected in increased instances of public-private partnerships [50].

Research indicates that collaborating with experts in different fields increases the actor 's knowledge base and ability to innovate by combining varying knowledge; this phenomenon is known as knowledge absorption which has been studied extensively within the pharmaceutical industry [51, 52]. However, this study contributes to the body of knowledge by observing this phenomenon within drug R&D.

Given that large pharmaceutical companies have resources to exploit comparative advantages of alliances, they are more likely to partner with actors that are relatively new players that complement their ability and are not yet embedded in the system [53]. These actors on the peripheries of the network include smaller biotechnology and pharmaceutical companies that have specialized

knowledge. By partnering with these outsider actors first, the large pharmaceutical companies have the ability to claim exclusive rights to their specialized knowledge before other actors. Therefore, we suspect that when large pharmaceutical companies collaborate or even acquire these companies, they are successfully gaining knowledge and resources that contribute to their ability to run successful trials that obtain regulatory approval.

One limitation to our study is that an approved drug may not necessarily be indicative to novel knowledge creation, since it may be a marginal improvement on an existing drug; there are many drugs that are considered "me-too drugs" which do not necessarily reflect a breakthrough in therapeutic effectiveness. Given the availability of data, we found it difficult to find a quantitative indication of true innovation. Although some studies use patents as indicators for innovation, the pharmaceutical industry files multiple patents for anyone compound that may have higher marketing potential in order to protect them from generic drug competition [54]. Nevertheless, we can still argue that a company is developing knowledge by accumulating clinical research data, personnel, and patient-base.

2.5 Conclusion

Our study contributes to the literature by empirical evidence that successful actors in the treatment development system will seek out diverse collaborators that are outsiders or new entrants into the market. As a result, successful actors tend to have lower cohesion in their local network. As the system evolves, we will inevitably notice a shift of knowledge concentration towards larger pharmaceutical companies, since there is evidence from this study and others that they are efficient at finding novel knowledge by searching for diverse partners. The role of outsider and new actors such as a start-up biotechnology or life science company will be crucial since the system is moving towards more cohesion, thus saturating the existing distribution of knowledge.

Applying the network analysis on traditional datasets allowed us to observe complex features that are not obvious using traditional measures such as firms and organizations. By observing the clinical trials as a system, we are able to examine the latent social structure that impacts the drug development process. We were able to calculate various network metrics on this large dataset. We further refine the research on pharmaceutical collaborations by focusing on the collaboration efforts within the clinical phases. Determining the causality was done by eliciting expert-opinions and synthesizing that knowledge into the analysis. Furthermore, we contributed by investigated various stakeholder

groups by characterizing them by actor types. Collaboration between firms using data regarding strategic alliances has been explored in the previous literature.

We can extend this work to include drug prices and its relations to the calculated metrics to observe the correlations between our current variables and drug prices. Another extension of this work could include analyzing the global metrics of the entire system and capturing the productivity of the system based on global network measures.

Chapter 3

Multi-component Integration of System Dynamics with Applications to World Population Growth and Sustainability

In this chapter, we present a multi-component model and the approach to the integration of the model. This large multi-component model will be presented in its entirety and proposed methods of integration. In the previous chapter, we examined how to handle complexity. Once complexity of the system is characterized quantitatively, we can proceed to model these behaviors as endogenous processes. We explore ways to build an integrated model that has

population dynamics as an endogenous variable in the model that is coupled within a feedback loop.

The intended purpose of this integrated model was four important questions: (i) how does human population growth impact climate change; (ii) what are the feedback loop effects between climate change and economic and social systems; (iii) what is the carrying capacity in terms of food, land, and natural resources and how does the population dynamics in different income regions behave due to an ecological overshoot; and (iv) how will inequality be exacerbated due to these changing factors and what kind of implications does that have on public health (e.g. disparity of food access during to famines).

This chapter aims at exploring how a large integrated model could be developed so calibration and testing are manageable. We present the framework of the integrated model with the relevant equations that correspond with the mechanics of each submodel, develop a plan of integrations of all model components, and propose and demonstrate an approach to calibrating each submodel.

3.1 Background

Understanding population growth will be more important than ever. It has been estimated that 9.7 billion people in 2050 and 11.2 billion people in 2100 by the United Nations [55]. The Intergovernmental Panel on Climate Change

(IPCC) business-as-usual projections predicting that global mean temperature will rise more than three degrees Celsius by the end of the century which will exacerbate issues such as drought and famines [56]. These issues are important and challenging to model since there are complex mechanisms at play.

In general, Bayesian population projection models [57, 58] developed by demographers and statistician have a strong validation but do not account for feedbacks. Other simpler models without feedback loops were developed as well to provide policymaker with an easy mathematical identity to provide policy-makers with a simple tool that is intuitive, such as the IPAT equation [59].

In 1971, the Club of Rome at the Massachusetts Institute of Technology commissioned the World 3 model in a publication called *Limits to Growth* [60]. This model was the first attempt at finding out the anthropogenic impacts and carrying capacity of the world using an integrated model that endogenized the population growth rate. However, the World 3 model was criticized as having parameters that were not based on existing social theories at the time. Furthermore, World 3 was accused of sensationalizing the existential threat of anthropogenic effects [61] (please refer to [62] for discussion). Other world models were developed to compete with World 3 [63–68].

In 1992, Integrated Assessment Models (IAM) were formally introduced

by William Nordhaus, an ardent critic of World 3, developed the Dynamic Integrated Climate Change Model (DICE) [69] which eventually won the Nobel Memorial Prize in Economic Sciences in 2018. DICE introduced a climate damage function and endogenized economic growth and climate change in a closed feedback loop. However, DICE does not endogenize population growth in their model. Other IAMs include [70–72] and also refer to a review of IAMs [73, 74]

Other models that were more theoretical explored growing inequality [75]. Models have also explored ways to integrate different submodels [76,77] within the main feedback loops.

3.2 Conceptual Framework

The goal of this model is to develop an integrated model of the global population dynamics with respect to future climate change and other sustainability trends. The conceptual framework of the model was based on causal diagrams that were developed from the expert consensus from a variety of disciplines in social sciences and public health. The model also investigates the growing economic gap between countries with low, middle, and high economic development that is documented in the literature. We will refer to the three collections of countries as income regions. For simplicity, we assumed that countries in each of the

three income group have identical attributes. Moreover, we also investigate the growing socioeconomic inequality within each region using economic theory regarding capital investment and economic growth rates.

The integrated dynamic model consists of seven different components or submodels that are coupled by feedback loops. This includes the following

- Population
- Health and Education
- Economy (Global and Regional)
- Global Natural Resources
- Global Climate System
- Global Water Resources
- Global Food System

The connections between each submodel are shown as arrows in Figure 3.1, and these arrows represent the causal links between the variables between submodels. In our model, we assumed that the population, health and education, and economy subsystems are regional since components will significantly differ between nations with different economic statuses. We modeled food consumption and production, water consumption, climate system, and energy consumption on a global scale because these resources are oftentimes traded between countries

with very different levels of development. Furthermore, we are interested in the relative depletion of resources on a global level resulting from different consumption from each income region. In order to illustrate the inequalities between the distribution of global reserve of resources, each income region in our model will have certain access to resources.

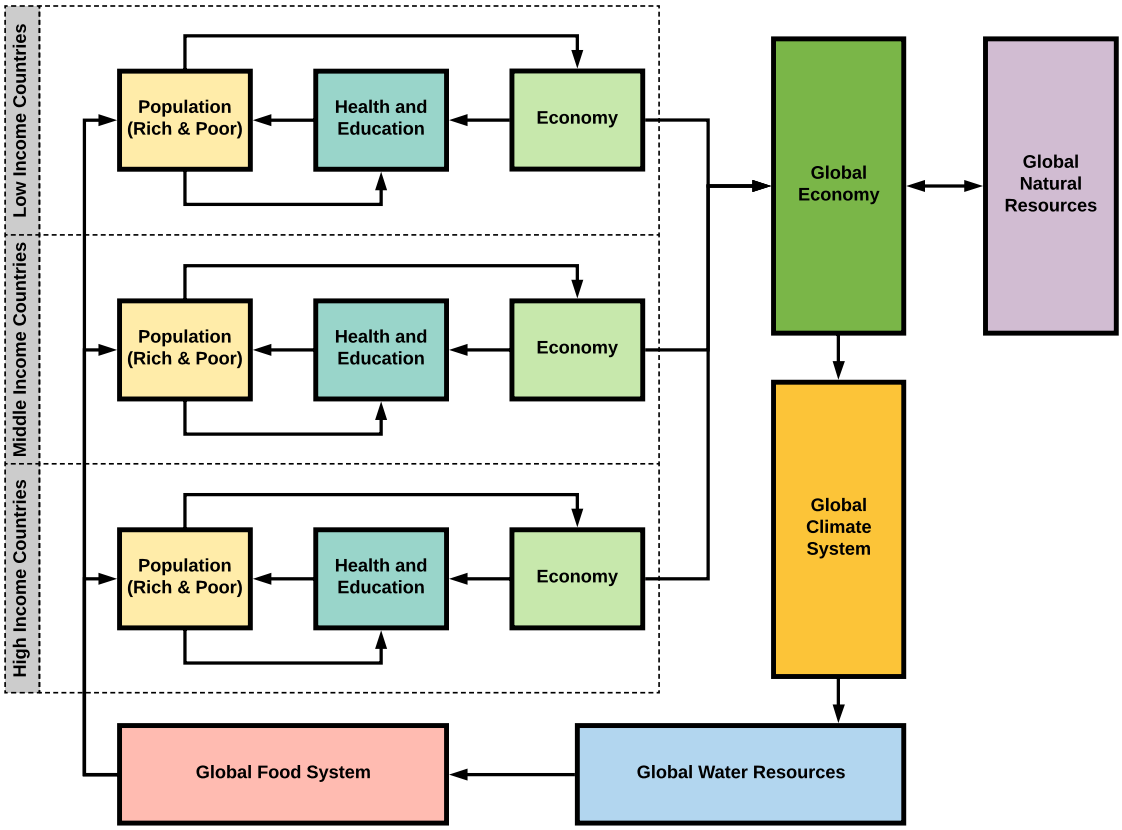


Figure 3.1: Overall Model Structure

For each income region, we divided the population into two socioeconomic groups: “rich” and “poor.” The access to health, education, and food for the two groups are different resulting in different fertility and mortality rates

of the population. As a result, we have two identical population subsystem for each socioeconomic groups in each income region. We defined poor as the poorest 20% of the population in terms of national income share. This allows us to distinguish the disparities in resource and service allocations that may impact the mortality and fertility rates of a population given some assumption or available data that distinguishes the different socioeconomic groups.

For the health and education submodels, we assumed that these social services are different for each income region. We also assumed that they are also the aggregate service volume of each region that within each region the health and education services are similar. The model also computes the aggregate economy in each region.

Afeter constructing the causal loop digram and high-level structure, we can fomulate this model based on the system dynamcis framework, which traditionally utilizes uses ordinary differential equations (ODE) such that **stock** variables (represented by $\mathbf{Y}$) are defined as

$$\frac{\mathrm{d}\mathbf{Y}(t)}{\mathrm{d}t} = \mathbf{X}^{\mathrm{IN}}(t) - \mathbf{X}^{\mathrm{OUT}}(t) \tag{3.1}$$

$$\mathbf{Y}(t) = \mathbf{Y}(t_0) + \int_{t_0}^{t} \mathbf{X}^{\mathrm{IN}}(\tau) - \mathbf{X}^{\mathrm{OUT}}(\tau)\,\mathrm{d}t \tag{3.2}$$

where x is the **inflow** and z is the **outflow**. In order to speed up our numerical computations, we decided to employ *difference equations* instead of continuous

differential equations to calculate our stock variables[1]. So,

$$\begin{aligned}\Delta \mathbf{Y}_{t+1} &\approx \frac{\mathrm{d}\mathbf{Y}(t)}{\mathrm{d}t} \\ &= \mathbf{Y}_{t+1} - \mathbf{Y}_t \\ &= \mathbf{X}_t^{\mathrm{IN}} - \mathbf{X}_t^{\mathrm{OUT}} \qquad (3.3)\end{aligned}$$

$$\therefore\ \mathbf{Y}_{t+1} = \mathbf{Y}_{t_0} + \mathbf{X}_t^{\mathrm{IN}} - \mathbf{X}_t^{\mathrm{OUT}} \qquad (3.4)$$

3.3 Submodels

3.3.1 Population (Regional)

As mentioned before, the population of each of the three income regions is divided into two socioeconomic subpopulations (rich and poor) with their own respective fertility and mortality rates. As a result, we have six subpopulations (3 regions $\times$ 2 socioeconomic subpopulation). The motivation for differentiating these groups is to show that certain changes in social services and food supply will affect the rate of change in different magnitudes between the two subpopulations in each income region as well as between the regions. For example, the shortage of food supply will affect the poorer demographics more since poor populations typically have less access to food than rich populations due to income constraints.

[1]However, for the sake of system dynamics convention, we will stick with the continuous mathematical definition of stocks and flows.

In each of the subpopulation, we modeled as a series of stocks and flows which depicts the population size of 8 age-sex groups. We defined four age-groups: 0-14 years, 15-49 years, 50-64 years, and 65+ years. We also assumed the traditional 2 sex groups: male and female. Figure 3.3 shows the causal relationships and stock and flows of each subpopulation. The stock and flow for each age-sex group are shown in Appendix C.1.

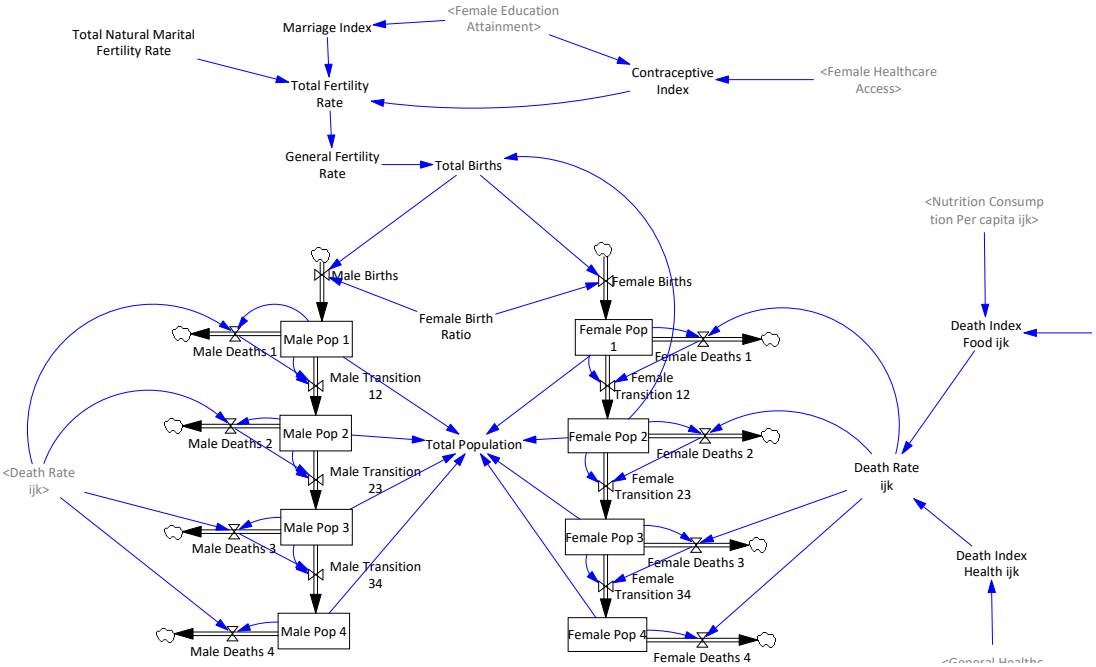


Figure 3.2: Population Subsystem

Total fertility rates are modeled based on Bongaart's proximate determinants of fertility, which is a foundational tool for demographers to project the total fertility rates TFR based on several indicators [78, 79]. The general form for

Bongaart's equation is

$$\begin{aligned} TFR(t) &= C_{\mathrm{M}}(t) \cdot C_{\mathrm{C}}(t) \cdot C_{\mathrm{A}} \cdot C_{\mathrm{I}} \cdot TF \\ &= C_{\mathrm{M}}(t) \cdot C_{\mathrm{C}}(t) \cdot TNM \end{aligned} \tag{3.5}$$

where C_{M}, C_{C}, C_{A}, C_{I} and TF represent the indices for **marriage**, **contraception**, **abortion**, **postpartum infecundity**, and **total fecundity**, respectively. The abortion index, postpartum infecundity index, and total fecundity TF are assumed to be fixed constants across time, which implies that the product of these three indices is constant. The product of these variables is referred as the **total natural marital fertility rate** (TNM) by [78]. We mathematically define this as $TNM = C_{\mathrm{A}}C_{\mathrm{I}}TF$.

Fertility rates are impacted by the level of female education access which is consistent with the literature [80], which directly impact C_{M} and C_{C}. We can also assume that female access to health services will impact the use of contraceptive C_{C} [81]. Hence,

$$C_{\mathrm{C}} = f(HF_{kr}, EF_{kr}) \tag{3.6}$$

$$C_{\mathrm{M}} = f(EF_{kr}) \tag{3.7}$$

The **mortality rate** DR_{ijkr} is associated with the **nutrition consumption**

per capita and **access to healthcare services**, defined as $\widehat{NC}_{kr}$ and HA_{kr} [82]. The impact on mortality rates varies depending on age, socioeconomic status, and level development on the country and will be estimated as

$$DR_{ijkr} = f(HA_{kr}, \widehat{NC}_{kr}). \tag{3.8}$$

Please refer to Section C.1 in Appendix C for complete listing of equations for the population submodel.

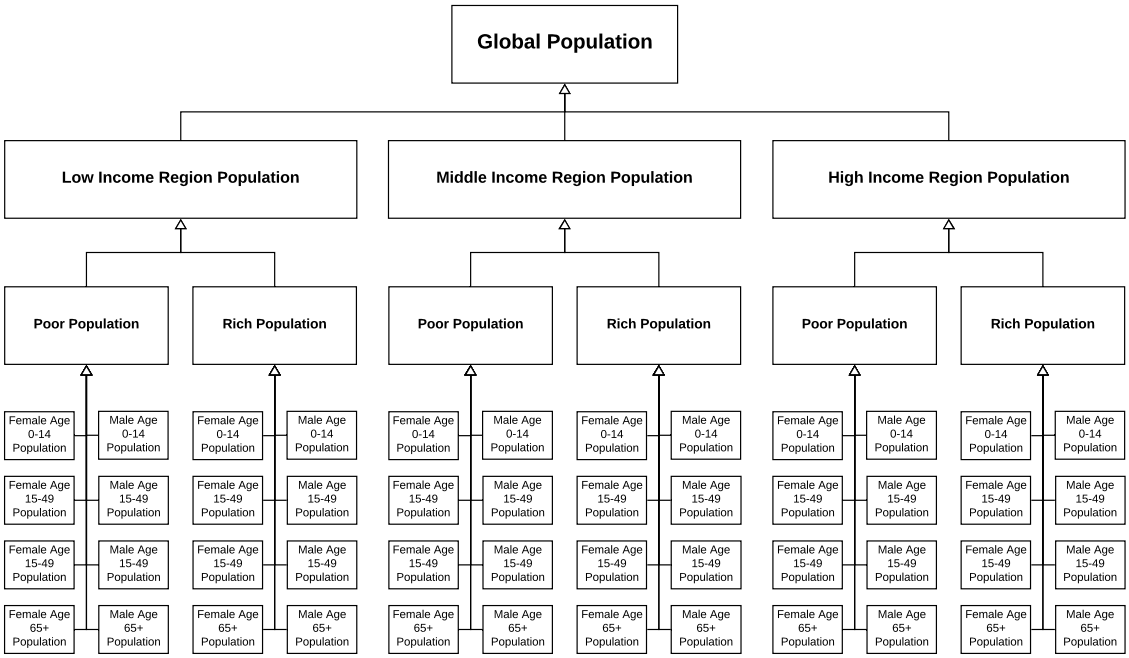


Figure 3.3: Population classification

3.3.2 Health and Education (Regional)

Health and education services are relevant to modeling population dynamics because they represent direct impacts to fertility and mortality rates. In the health and education submodel for each region, we assumed that there are three different levels of health and education systems in the world that corresponds with the three income region, which means that all countries in a specific region receive the same level of services. Since we are aggregating the population, it is fair that the services should be homogeneous.

Health services and **education services** are defined as stocks which represent the volume of services in their respective sector (see Figure 3.4). We quantify this service as the volume of investment that is allocated from the total economic output. Health services are supposed to represent the number of physicians, hospitals, and health care systems in a country. The stock of education services represents the number of schools, universities, and educational professions in a country.

Both services have a fixed rate of depreciation that is a function of the stock size. We assumed that the larger the amount of services available, the higher the depreciation outflow is. As a result, the regions have to invest economic resources into these subsystems to maintain operations. If the volume of the economic resource exceeds the depreciation threshold, then the stock of services will grow, and vice versa.

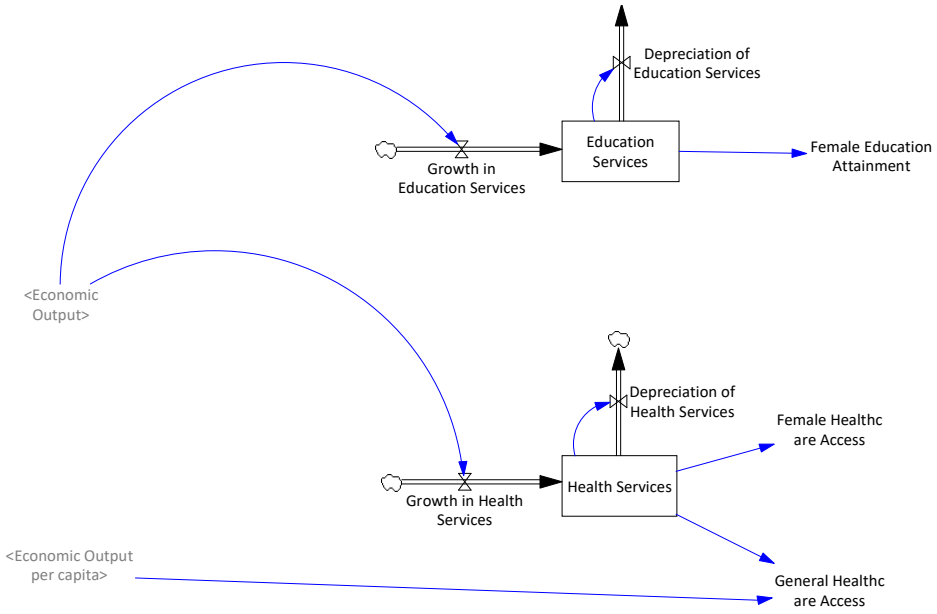


Figure 3.4: Health and Education Submodel

The stock of education directly influences female education attainment which can be thought of as the average level of education of women. Similarly, the stocks of health and education services influence the female healthcare access that relates to family planning and contraceptive services and general health care access of the population. The variables, female education attainment, female healthcare access, and general health care access, all feed into the population submodel.

The equations are located in Section C.2 in Appendix C.

3.3.3 Economy (Regional and Global)

We modeled the economy based the three income regions with the Solow growth model. that utilizes four input factors: labor, capital, nonrenewable resources, and renewable resources to calculate the economic output of each region. The **economic output** Y is a Cobb-Douglas production function explicitly defined as

$$Y(t) = A(t) \cdot L(t)^{\eta_1} \cdot K(t)^{\eta_2} \cdot N(t)^{\eta_3} \cdot R(t)^{\eta_4} \tag{3.9}$$

The **technology multiplier** which is also know as the total factor productivity A is assumed to be time variant, and can be modeled as a linear function to demonstrate exogenous growth with respect to time. The **labor**, **capital**, **nonrenewable resources**, and **renewable resources** are input factors, L and K, have corresponding **input production elasticities** η_1, η_2, η_3, and η_4. For simplicity, we will assume constant returns to scale, or $\eta_1 + \eta_2 + \eta_3 + \eta_4 = 1$ for the four input factors. Refer to Appendix C.3 to see how we incorporated the production function. Typically, the economic output is represented as the Gross Domestic Product (GDP).

We assumed that the **employment-to-working population ratio** is fixed since we are interested in long-term trends. However, the employed labor force will adjust based on the population size of the working-age groups, which is defined as the two age-groups between 15-64 years-old. Capital is defined

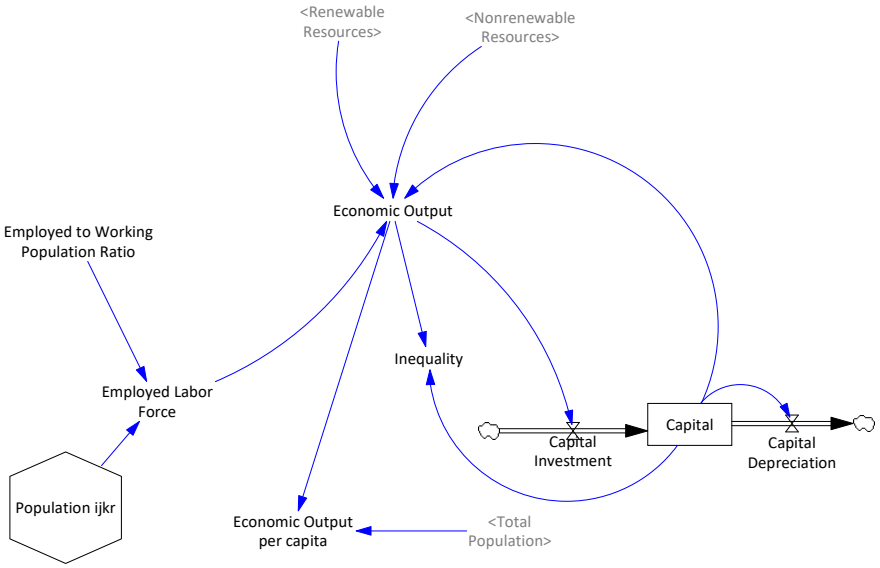


Figure 3.5: Economic Submodel

in the neoclassical sense in which it includes the real value of machinery, equipment, and infrastructure. Nonrenewable and renewable resources are inputs into the economic submodel from the resource submodel. There is a feedback loop between the factor inputs and consumption rate. Mathematically, we used the stock levels of nonrenewable and renewable resources to calculate economic output instead of the resource consumption in order to provide a time-delay for consumption of resources.

Inequality is modeled based on the conclusion of Thomas Piketty [83] that the return on capital versus the economic output growth rate is causing socioeconomic inequality since a fraction of capital ownership belongs to the wealthy. As a

result, we can mathematically express inequality, Q, as

$$Q(t) = \Psi \cdot \frac{\Delta K(t)}{\Delta Y(t)}. \tag{3.10}$$

We used **inequality** to determine the poor/rich access to food and healthcare in each region. Inequality is defined as the share of national income for the poorest quintile of the population. We define "poor" as the poorest quintile of the population and the rest as "rich." We can figure the income per capita for poor and rich groups based on the Q measure.

For a complete listing of equations in the economic submodel, refer to Section C.3 in Appendix C.

3.3.4 Global Natural Resources

The natural resources are divided into two types: renewable and nonrenewable resources. We define nonrenewable resources like fossil fuels and renewable resources as biomass and timber. It is assumed that the consumption of these resources is caused by the combined economic activities of Low, Middle, and High-income regions.

In Figure 3.6, we can see that the consumption of non-renewable and renewable resources are modeled as outflows. We define economic activities as any activity that contributes to the Gross Domestic Product (e.g. energy usage, manufacturing,

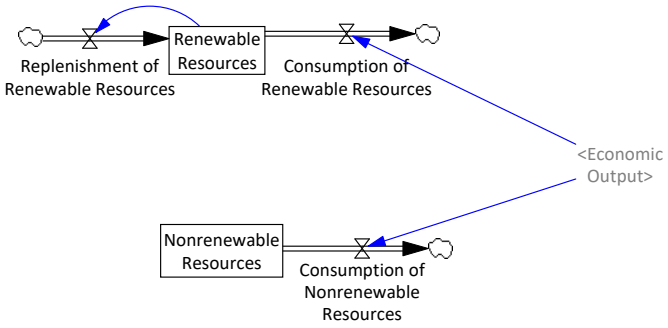


Figure 3.6: Global Natural Resources

mining). There is a balancing feedback between renewable/nonrenewable stocks and economic growth because these resources also serve as production input factors for the Solow Growth Model shown in Equation (3.9). The stocks are meant to represent the reserves of these natural resources in the world. Renewable resources by definition have a replenishment rate, which is shown as an inflow in Figure 3.6.

For a complete listing of equations in the resources submodel, refer to Section C.4 in Appendix C.

3.3.5 Global Climate System

The global climate system is represented by a simplified model of global temperature and CO_2 concentration in the atmosphere. CO_2 concentration increases the radiative forcing in the atmosphere. We assumed the other greenhouse gases other than CO_2 are also contributing to the radiative forcing effects that

drive temperature change. However, this rate is currently constant since we assume that the volume of economic activities corresponds to the amount of CO_2.

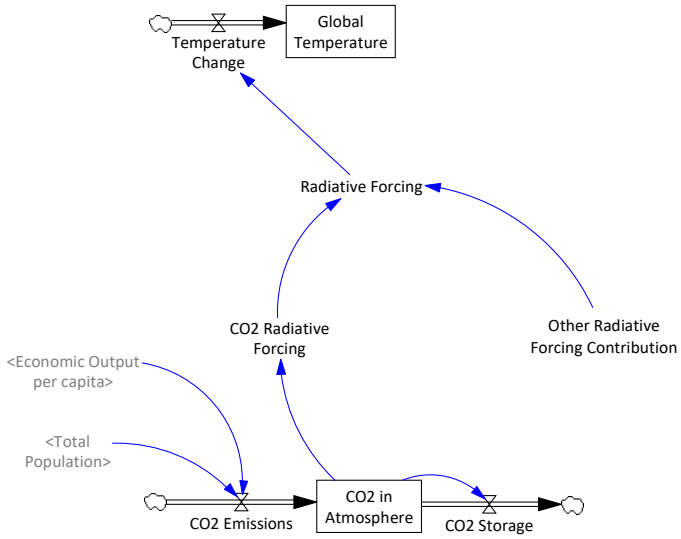


Figure 3.7: Climate Change Submodel

We can define climate change by using the IPCC equation that relates radiative forcing RF with changes in equilibrium surface temperature [56]. The exogenous variable λ is the climate sensitivity, which is determined from previous studies [84].

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = \lambda \cdot RF(t) \tag{3.11}$$

The radiative forcing is calculated using a well-known equation that determines

the effects of CO_2 concentrations in the atmosphere [85]. This relationship is defined in Equation (3.12) where G and G_0 are respectively the concentration and reference concentration[2] of CO_2. Radiative forcing from other greenhouse gases (GHG) is considered by the exogenous parameter RF_{EXT}.

$$RF(t) = 5.35 \cdot \ln \frac{G(t)}{G_0} + RF_{\mathrm{EXT}} \tag{3.12}$$

For a complete listing of equations in the climate system submodel, refer to Section C.5 in Appendix C.

3.3.6 Global Water Resources

The water resources are modeled as a pooled resource for the whole world. We assumed that freshwater supply refers to the amount of water accessible for utilization (e.g. aquifers, lakes). In the global water resources submodel, we assumed that the reserve of freshwater is a renewable stock with a replenishment rate. The replenishment rate in our model is assumed as a fixed proportion of the freshwater stock.

Based on research involving the variability of precipitation due to climate change, it has been estimated that there could be a net increase in population that experiences water stresses [56, 86–89]. As a result, we have added an

[2]The reference concentration of CO_2 is commonly equal to the baseline levels of CO_2 that is measured between 1951-1980. This is the standard that NASA uses since these measurements are determined to be the earliest, most accurate values.

outflow to the freshwater stock that represents the proportion of the freshwater stock that is lost due to climate change effects. We have decided that the water loss due to climate change will be diminishing in relations to temperature. As a result, we modeled the water loss WL due to climate change as

$$WL(t) = W(t) \cdot CC_{\mathrm{W}}(t) = W(t) \cdot \left(\delta_{\mathrm{W}} \cdot \frac{T_0}{T(t)} \right) \tag{3.13}$$

where CC_W represents the proportion of water stocks lost to climate change effects. This proportion is assumed to be a ratio of the baseline temperature T versus temperature T at time t with a multiplicative coefficient δ_{W}.

We determined that there are three primary uses for water: municipal, agricultural, and industrial uses. Municipal use refers to water being utilized by households for average living and drinking purposes. Agricultural uses of water refer to water utilized in food production of meat and crops. We assume that the industrial use of water is driven by economic activity. In the current version, we are only simulating the water shortage effects with regards to the food supply.

For a complete listing of equations in the water submodel, refer to Section C.6 in Appendix C.

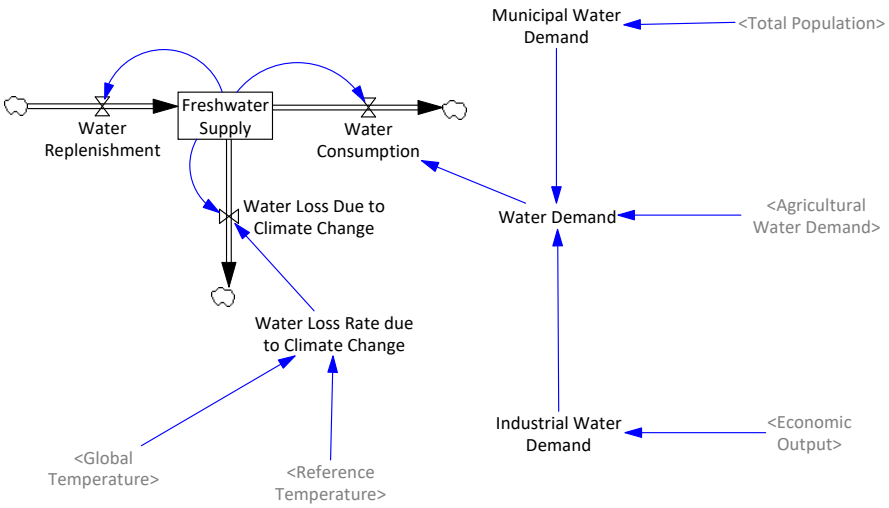


Figure 3.8: Water Submodel

3.3.7 Global Food System

In order to simplify the food system so we captured international trade of food commodities, we examined the food system on a global-level and assumed that the production of food is dictated by the global demand and supply. We consider three types of food: fish, livestock, and crops. The inputs factors of food productions, grazeland, cropland, fisheries, and water, are also modeled as globally pooled resources.

To consider the disparity between income regions as well as socioeconomic strata we utilized a partition function that determines the access to food resources for each income regions and the rich and poor populations of the region.

In order to model the varying demand/production of food versus income,

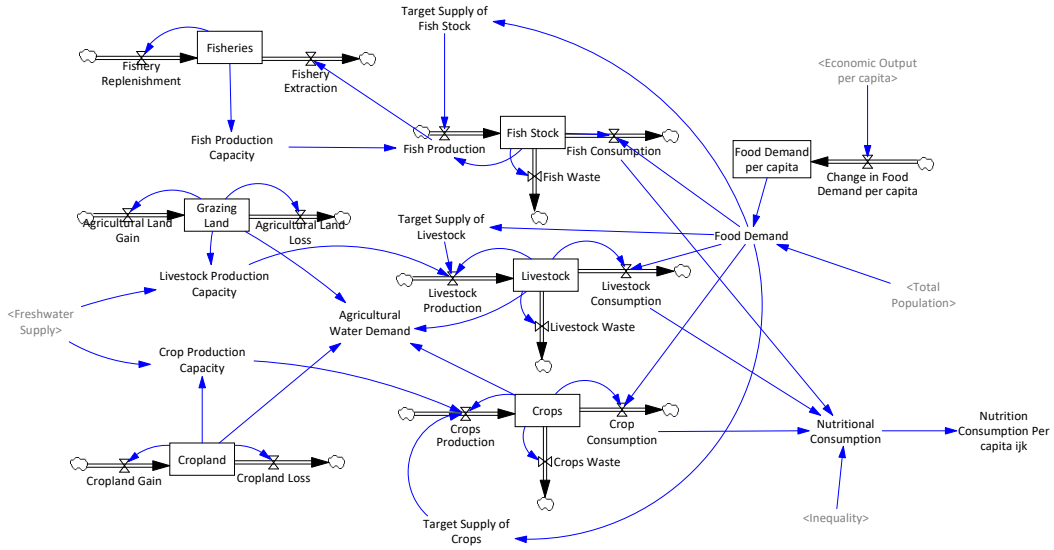


Figure 3.9: Food Submodel

we utilized the microeconomic concept of *income elasticity of demand*. The income elasticity of demand measures the magnitude to which food demand changes relative to the change in income of a person. We have assumed there are three different food demand per capita based on the income region since the demand would be different given higher income regions would have less growth in demand given an increase in income, and vice versa. This phenomenon is known as Engel's Law and has been widely observed [90]. Food demand per capita is modeled as stocks (shown in Figure 3.9) since we are looking at first-order changes between income per capita.

For a complete listing of equations in the food submodel, refer to Section C.7 in Appendix C.

3.4 Data Collection and Processing

In order to estimate the parameters, we must find appropriate indicators to represent the state variables in our model. We defined state variables as the output variables that we are observing. These variables vary in units, therefore, it is important for us to verify the consistency in unit conversions.

We utilized several data sources to inform the model and parameterize the model coefficients. The data sources ranged from the following.

- Demographic Health Survey, USAID
- World Development Indicators, The World Bank
- International Labour Organization Statistics Database
- United Nations Population Statistics
- World Income Inequality Database, United Nations University
- Global Health Expenditure Database, World Health Organization
- BP Statistical Review of World Energy

3.4.1 Population Submodel Data

Using the World Bank's classifications of countries into income groups, we were able to adopt these regions for our model. Please make note that the

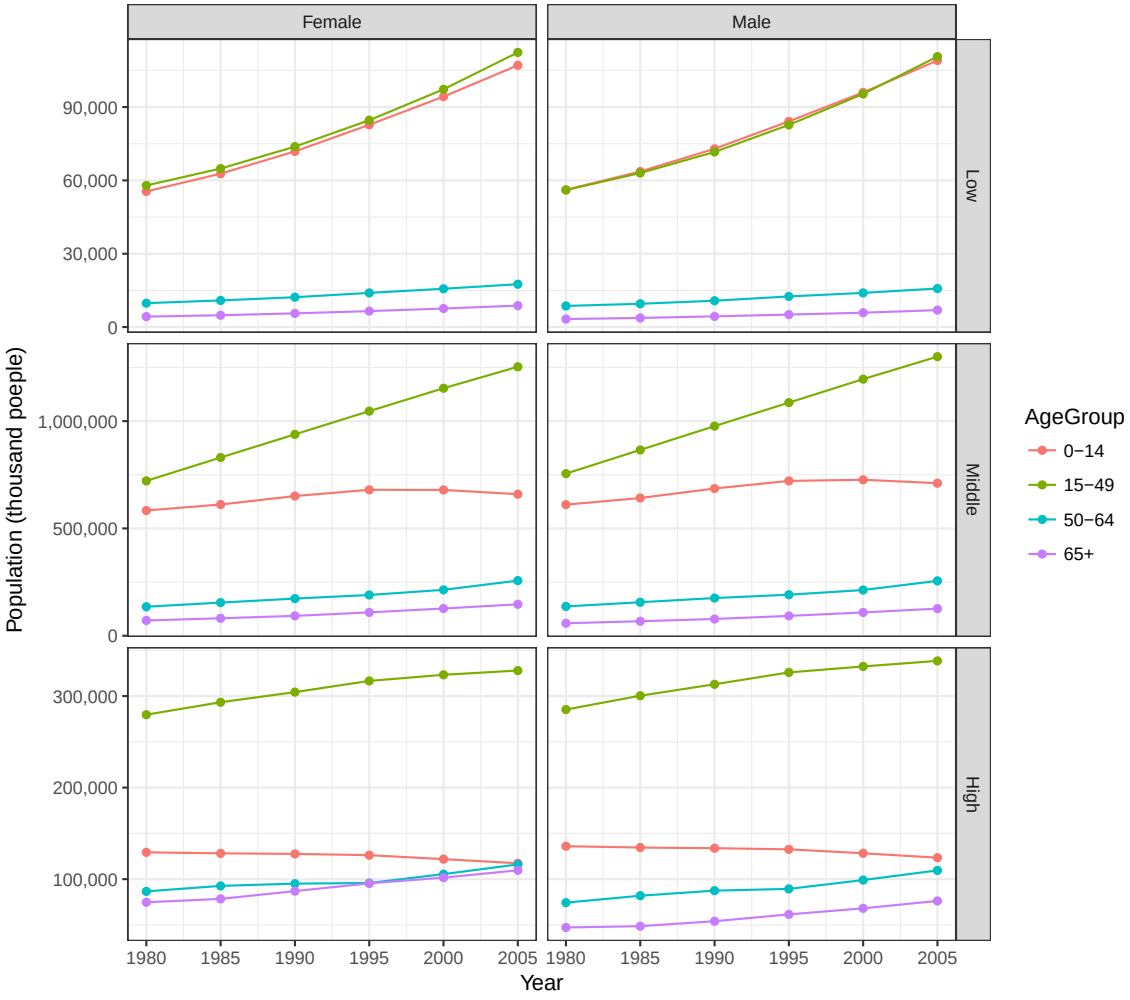


Figure 3.10: Population data for each income region from 1990-2005 based on the United Nations Population Statistics. We use this information to calibrate the model.

World Bank has 4 regions: low, middle (lower-middle + upper-middle), and high income – we decided to combine lower-middle and upper-middle income groups as one income region. The population for each income region is calculated by simply summing the population of all the countries in that particular region. The population size input data for each region are plotted in Figure 3.10.

3.4.1.1 Estimating Bongaart's Proximate Determinants

Let's recall Bongaart's equation that models the proximate determinants of fertility which is shown below and in Equation (3.5). These two proximate determinants are composed of the marriage index $C_{\mathrm{M}kr}$ and contraceptive index $C_{\mathrm{C}kr}$.

$$TFR_{kr}(t) = C_{\mathrm{M}kr}(t) \cdot C_{\mathrm{C}kr}(t) \cdot TNM$$

Based on existing literature, we were able to assume that the "level of women education attainment" influences the fertility determinant, marriage ($C_{\mathrm{M}kr}$). Likewise, education attainment and "women's access to healthcare" is also a factor that influences the prevalence of contraceptive use ($C_{\mathrm{C}kr}$).

Women education attainment is quantified by As for contraceptive usage ($C_{\mathrm{M}kr}$),

As mentioned before, we looked at education attainment EF_{kr} and female health access HF_{kr} as the primary drivers that controls the marriage index $C_{\mathrm{M}kr}$ and contraceptive use $C_{\mathrm{C}kr}$. We assumed a linear functional relationship (see Equations (3.14) and (3.15)) such that the coefficients α_{M}, α_{C}, β_{EM}, β_{EC},

and β_{HC} can be estimated using a linear regression.

$$C_{\mathrm{M}kr}(t) = \alpha_{\mathrm{M}} + \beta_{\mathrm{EM}} \cdot \big(1 - EF_{kr}(t)\big) \tag{3.14}$$

$$C_{\mathrm{C}kr}(t) = \alpha_{\mathrm{C}} + \beta_{\mathrm{EC}} \cdot \big(1 - EF_{kr}(t)\big) + \beta_{\mathrm{HC}} \cdot \big(1 - HF_{kr}(t)\big) \tag{3.15}$$

In our linear regression, we used country-level data for the input variables, EF_{kr} and HF_{kr} to predict marriage $C_{\mathrm{C}kr}$ and contraceptive-use rate $C_{\mathrm{C}kr}$. The data that is used to estimate these parameters are extracted from the Demographic Health Survey (DHS) conducted by USAID.

For our regression analysis, we used *marital status of women (Married or living in union)* to represent $C_{\mathrm{F}kr}$ and *current use of any modern method of contraception (all women)* as $C_{\mathrm{C}kr}$. The independent variable $EA_{\mathrm{F}kr}$ is informed by *women with completed secondary education*, and $HA_{\mathrm{F}kr}$ by the *percentage of women that had their delivery at a health facility (3 years before the survey)*. We used 239 observations to estimate $C_{\mathrm{M}kr}$ and 222 observations to estimate $C_{\mathrm{C}kr}$. For both regressions, there were 222 observations spanning 74 countries and 26 years of surveys from 1990 through 2016. The data and results of the regression are listed in Table 3.1.

Coefficient	Estimate	SE	T-stat	P-value
α_{M}	0.4514	0.0612	7.3751	$<$0.001
β_{EM}	0.2022	0.0651	3.1028	0.002
α_{C}	0.4141	0.0657	6.2956	$<$0.001
β_{HC}	0.1538	0.0407	3.7753	$<$0.001
β_{EC}	0.3119	0.0805	3.8714	$<$0.001

Table 3.1: Regression Coefficient Values

3.4.1.2 Estimating Lorenz Curve and distinguishing rich and poor

Using the data from the World Income Inequality Database from the United Nations University, we were able to reconstruct the Lorenz curve for each region. The Lorenz curve is used to determine the population size of the two socioeconomic groups: rich and poor. We were able to calculate the Lorenz curve for each income region by aggregating the socioeconomic groups for each region based on income share percentage. The aggregated Lorenz curve for each region is calculated for each ten-year interval with the results being shown in Figure 3.11. We used these values to figure out the distribution of the population in rich and poor socioeconomic groups. We considered the split between poor and rich population to be the 20 cumulative percentile of income share.

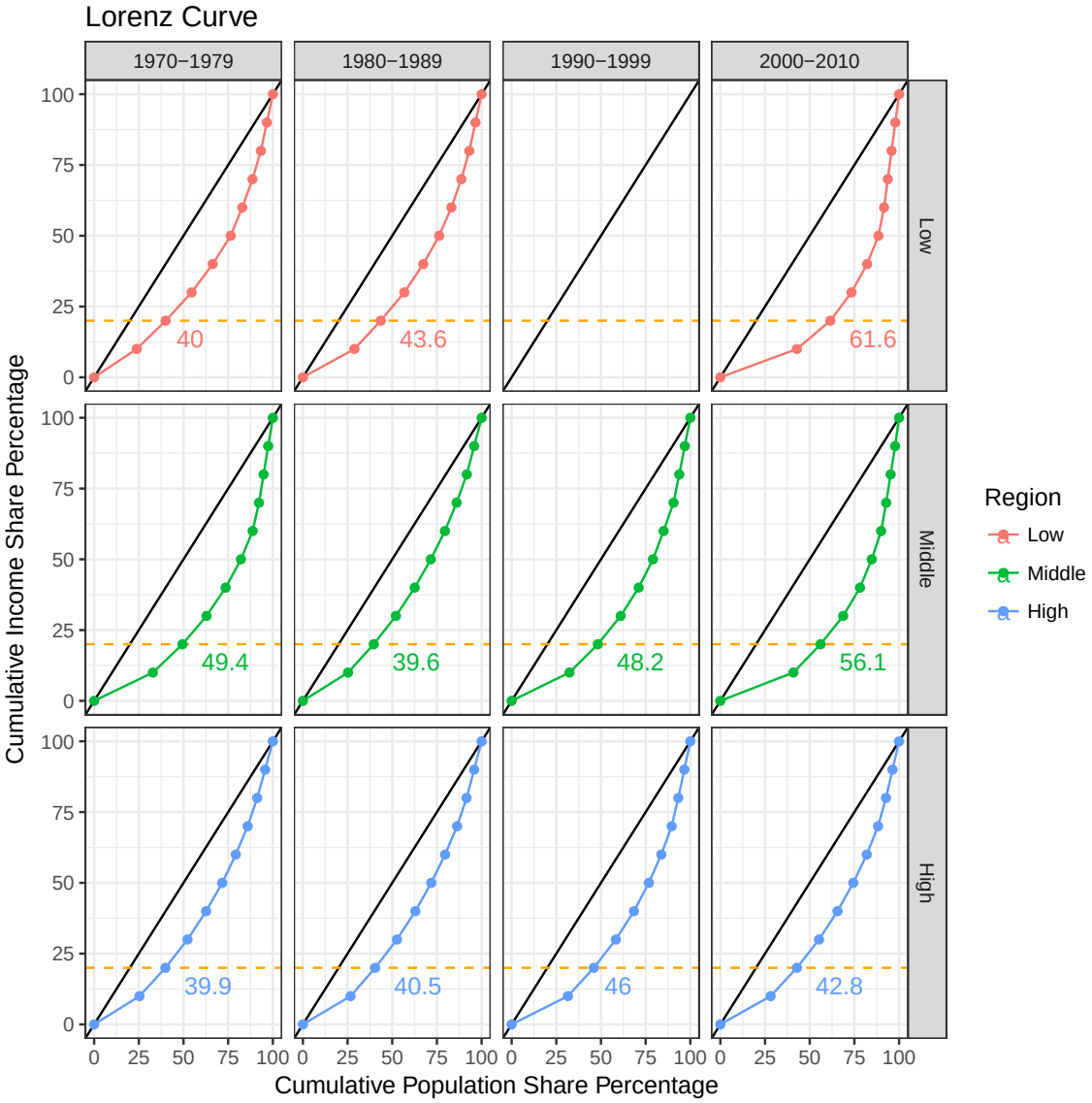


Figure 3.11: Aggregated Lorenz Curve for each income region is shown. The bottom percentage of the population that has 20 percent of the regional income is indicated by the orange line. The intersection value is shown by the number on the right side of the Lorenz curve for each period. We utilized this information to determine the population size of the two socioeconomic groups.

3.4.2 Health and Education

The stock of health and education services are impacted by variables directly, cost and investment. We assumed that a fixed percentage of economic output (i.e. GDP) in each region goes towards health and educational expenditures, which we modeled, respectively, as stocks of health and educational services. Based on the World Health Organization (WHO) and the World Bank, we were able to estimate these stocks using the expenditures volume.

The health-related variables in the health and education submodel consist of 3 output variables that are parameterized based on empirical data from the World Health Organization (WHO): **health services** (stock variable), **female healthcare access** (auxiliary variable), and **general healthcare access** (auxiliary variable).

To be consistent with the population submodel, we will use *percentage of children that had their delivery at a health facility (3 years before the survey)* to represent female healthcare access for each income region. We determined the income region's health access percentage as the average of all the countries that belong to that income region. This aggregation method is conducted for each 5-year interval ranging from 1985-2015. We utilized the percentage of children that had their delivery at a health facility to as a surrogate measure for the "rich" socioeconomic group in each region. Additionally, we were given information on female access to health care for the *percentage of children in the*

poorest 20 percent of the population that had their delivery at a health facility (3 years before the survey), which we used to calculate the health care access for the "poor" socioeconomic group.

To calibrate the stock of health services, we decided to use *health capital expenditures* as a representation for the amount health services (e.g. hospital, infrastructure) for each income region that is provided by the WHO's Global Health Expenditure Database. The health capital expenditures for each income regions is calculated by aggregating all corresponding countries for each year ranging from 2000-2015.

We defined general access to healthcare as the ability of the general population to receive quality healthcare services. As a result, we utilized the *Healthcare Access and Quality Index* (HAQI) developed by the Global Burden of Disease Collaborative Network and downloaded from the Institute for Health Metrics and Evaluation [82]. The average value of the HAQI measure is used to calibrate the general health care access for the rich socioeconomic group while the lower bound of the 95% uncertainty interval is used to calibrate general health care access for the poor socioeconomic group. This aggregation is done for every 5 year-interval ranging 1990-2015.

There are two variables that we calibrated in this subsystem includes **education services** and **female access to education**. The UNESCO Institute for Statistics (UIS) database provided information on *government expenditures on education*

which we used as a surrogate for educational services. Since the availability of data for the time-series is sparse, we took the mean of government expenditures on education for each year ranging from 2011-2014.

Consistent with our regression estimating Bongaart's Proximate Determinants ((3.15) and (3.14)), we used the *percentage of female population age 15+ that completed secondary education* obtained from the Barro-Lee Educational Attainment Dataset.

3.4.3 Economy

Utilizing data from the World Bank, we were able to directly extract the aggregate *Gross Domestic Product (GDP)* of each income region (Low, Middle, and High-income countries) for each year from 1980 to 2015 to determine **economic output**. **Capital** stock data is extracted from *Public and Private Capital Investment* category the Investment and Capital Dataset compiled by the International Monetary Fund (IMF). We aggregated public and private capital in our analysis as the total capital investment. The data is used to compute the **factor productivity elasticities** in our Solow Growth model formulation for each income region. *Workforce participation rate* is collected from the International Labor Organization (ILO) Database which allowed us to calculate the labor size **Labor** from population size data.

3.4.4 Resources

We collected the data regarding nonrenewable natural resources that are required for energy production based on BP Statistical Review of World Energy Dataset. From this dataset, we were able to extract *Oil, Coal, and Natural Gas Consumption*, as well as *Oil, Coal, and Natural Gas reserves in tonnes*, which we will assume to be **nonrenewable resource stock**.

The **renewable resource stock** is assumed to be biomass such as timber cover. This is extracted from the National Footprint Accounts 2018 Public Data hosted by the Global Footprint Network. We will use the *forest cover area* (acres) as the unit of measurement.

3.4.5 Climate

Data regarding **mean global temperature** is collected from the US National Aeronautics and Space Administration (NASA). The World Bank provides data on the $\mathbf{CO_2}$ **emission per capita** based on income regions.

3.4.6 Water

We used **global water reserve** estimates from [91]. The study also measures the **global water consumption** as well as **global water replenishment rate**.

3.4.7 Food

Using the Food and Agricultural Organization Statistics Database (FAOSTAT), we were able to calculate the stock for the 3 food types: **Fishstock**, **livestock**, and **crops** based on the *food supply in tonnes* at each year. Food **consumption**, **production**, **nutrition consumption**, and **loss** were also extracted from FAO database. Based on the this information, we were able to calculate the global consumption as well as the regional consumption since the the data was classified into separate countries.

3.4.8 Data Sources

Table 3.2: Data Sources

Submodel	State Variable	Measurement	Data Source
Population	Population size P_{ijkr}	People	UN Population Division
Health & Education	Health services HS_r	Health Capital Expenditures (current USD)	WHO Global Health Expenditures Database
Health & Education	Female healthcare access $HF_{\mathrm{F}kr}$	Percentage of children that had their delivery at a health facility (3 years before survey)	UNICEF Maternal and Newborn Health Coverage Database
Health & Education	General healthcare access $HA_{\mathrm{F}kr}$	Healthcare access and quality index (HAQI)	Institute for Health Metrics and Evaluation
Health & Education	Education services ES_r	government expenditures on education (current USD)	UNESCO Institute for Statistics (UIS) Database
Health & Education	Female education attainment EF_r	percentage of female population age 15+ that completed secondary education	Barro-Lee Educational Attainment Dataset
Economic	Inequality Q_r	Bottom percentage of the population that shares 20% of national income	UNU World Income Inequality Database
Economic	Economic Output Y_r	Gross Domestic Product (current USD)	World Bank Databank
Economic	Capital K_r	Public and Private Capital Investment (current USD)	IMF Investment and Capital Dataset
Economic	Labor L_r	Labor Force Participation Rate (current USD)	ILOSTAT Dataset
Resources	Nonrenewable Resources N	Coal, Natural Gas, and Oil (tonnes and barrels)	BP Statistics Dataset
Resources	Renewable Resources R	Forest Cover (acre)	GFN National Footprint Accounts 2018 Public Data
Climate	Mean global temperature T	Mean global temperature (Celsius)	NASA Open Data
Climate	CO_2 Emission per capita $\widehat{G}b$	CO_2 emission per capita (ton per capita per year)	World Bank Databank
Food	Livestock Production FP_{MEAT}	Livestock Production (tonnes)	FAOSTAT Commodity Balance
Food	Fish Production FP_{FISH}	Fish Production (tonnes)	FAOSTAT Commodity Balance
Food	Crop Production FP_{CROP}	Crop Production (tonnes)	FAOSTAT Commodity Balance
Food	Livestock Consumption FC_{MEAT}	Livestock Consumption (tonnes)	FAOSTAT Commodity Balance
Food	Fish Consumption FC_{FISH}	Fish Consumption (tonnes)	FAOSTAT Commodity Balance
Food	Crop Consumption FC_{CROP}	Crop Consumption (tonnes)	FAOSTAT Commodity Balance

3.5 Model Integration

The overall structure of the model is divided into a multi-component model in R. The main relationships between each submodel outlined in the previous section are explicitly structured in the R code as shown in Figure 3.13, such that a submodel is represented as a separate component of the model. Specifically, each submodel is defined as a separate function in R.

Certain regional submodels have region-specific outputs that are combined by *aggregation functions* to become one global-scale variable that acts as a component input into global-scale submodels. On the other hand, some of the global-scale components have global-scale outputs that are reallocated to each region by a *partition function*. The modularity of each component allows us tremendous flexibility in increasing the complexity of mechanics in each submodel, as well as the expansion of the model in future iterations.

In order to speed up the simulation, all differential equations were converted to difference equations. In essence, we assumed

$$\frac{\mathrm{d}\mathbf{Y}(t)}{\mathrm{d}t} \approx \mathbf{Y}(t+1) - \mathbf{Y}(t) \tag{3.16}$$

Each computational run is approximately one second. This much faster than using a numerical solver for solving ordinary differential equations. We assumed a time step of one year and ran the model from the year 1980 to 2080. The

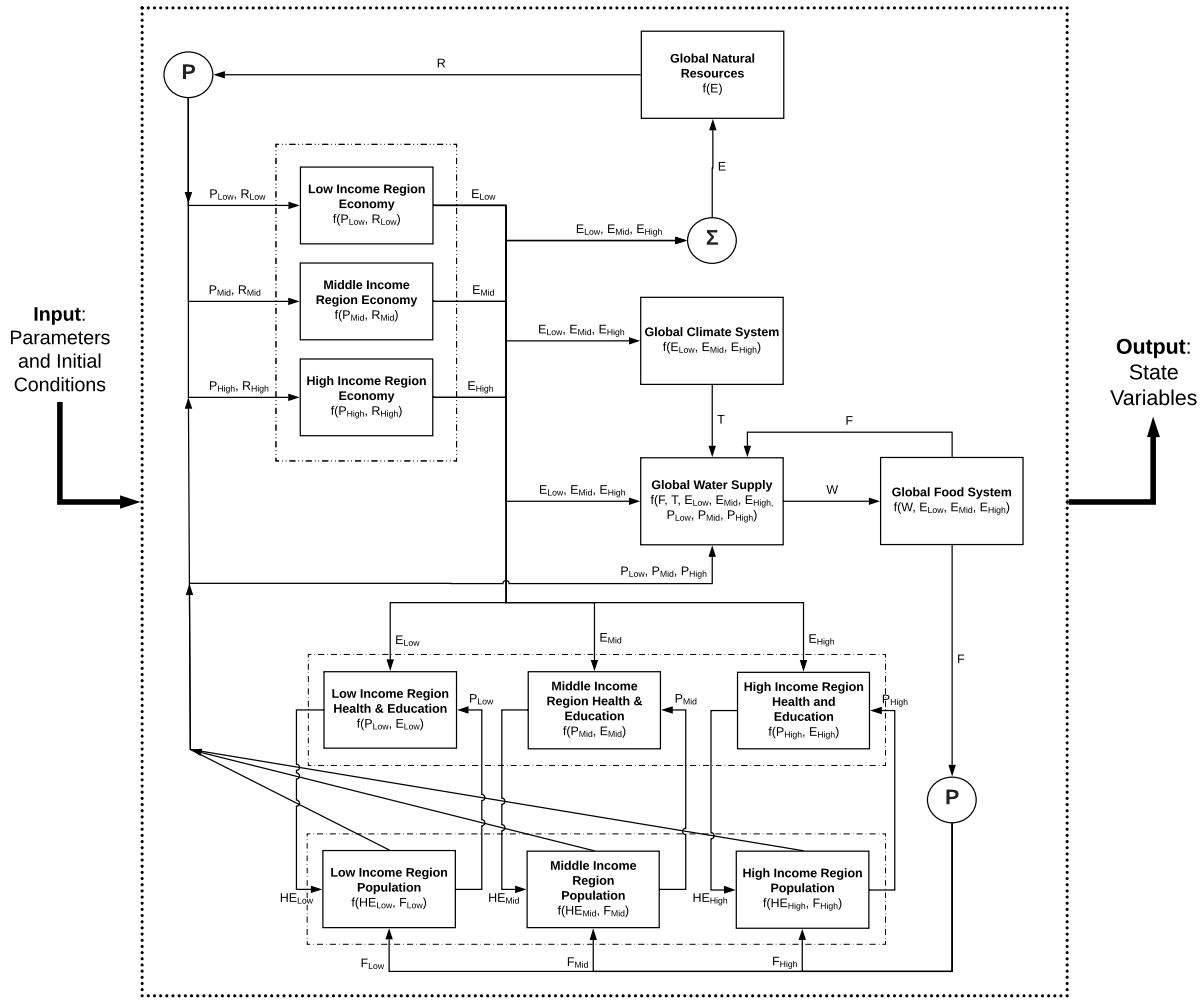


Figure 3.12: Component Diagram of the entire model. The aggregation functions are represented as Σ and the partition functions are represented as P. The submodels (component) of the model are depicted as solid squares. The input and output variables for each component are shown as the arrows. The variables are generalized as P_k = population size, E_k = economic output, F_k = food/nutrition consumption, R_k = nonrenewable/renewable resources, HE_k = health and education access, and T = global temperature.

initial conditions were informed by external data sources which were described

in Section 3.4.

3.6 Submodel Parameter Estimation

In order to solve our problem, we utlized the *least-squares ftting* to estimate our parameters based on points from 1980 to 2015, which were collected and outlined in Section 3.4. Lets say we have N number of state variables (output variables) in our model and T number of observed poionts. We have empirical data with one realization for each variable y_{it} and can simulate the data using the integrated model $\widehat{y}_{it}$. We can then define a cost function as the sum of square errors Err with weights w_i for each variable i.

$$Err = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{y_{it} - \widehat{y}_{it}}{w_i} \right)^2 \tag{3.17}$$

We can generalized our integrated nonlinear model as

$$\widehat{y}_{it} = f(x_{ijt}, a_{ij}) \tag{3.18}$$

where x_{ijt} are system input variables a_{ij} are coefficients corresponding each system variables (assuming there is a coefficient corresponding with each input variable).

We also know the minimum (a_{ij}^{MIN}) and maximum possible values ($a_{ij}^{\max}$) of each coefficient values. This allow us to estimate the parameters using the

following constrained-optimization problem.

$$\min_{a_{ij}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\frac{y_{it} - f(x_{ijt}, a_{ijt})}{w_i} \right)^2$$

$$\text{s.t.} \quad a_{ij}^{\min} \leq a_{ij} \leq a_{ij}^{\max} \tag{3.19}$$

In order to reduce the number of parameters that need to be calibrated by optimization, we have elected to calibrate the submodel individually. We assumed the nonlinear least-squares problem (3.19). The parameterization was carried out using the `GA` package in R. The `GA` library includes several metaheuristic genetic algorithm solvers that allow the user to easily define the cost function (error function) and the parameters.

Based on the structure of our model, the submodel components are able to be isolated such that the parameters in each submodel are solved individually with *endogenous inputs* into each submodel being substituted by *exogenous inputs*. We demonstrate the parameter estimation process in Figure 3.13.

We utilized the *gaisl* function in the library `GA` to calibrate the coefficients and fit our difference equation model. The computation was conducted using parallel computing with Islands Genetic Algorithms meta-heuristic developed by [92].

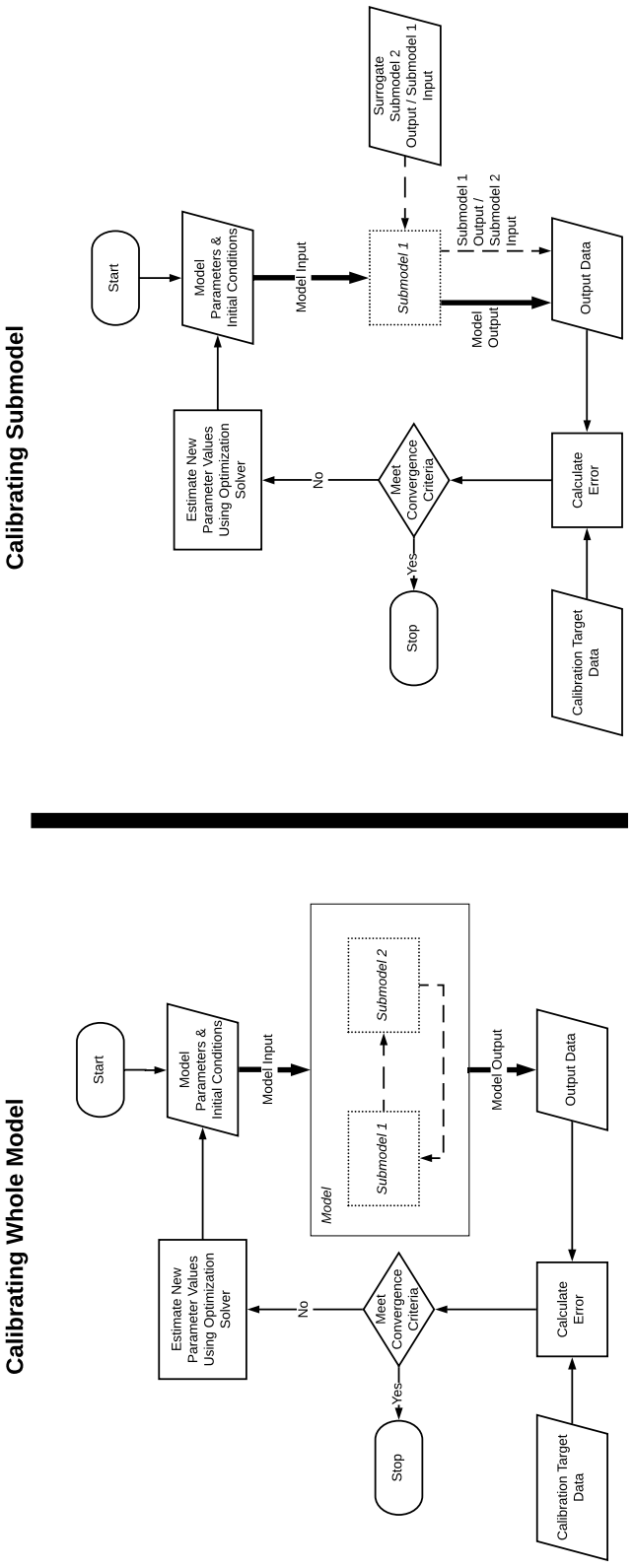


Figure 3.13: Whole Model versus Submodel Calibration Flow Chart.

3.6.1 Calibrating Health and Education Submodel

In this section, we demonstrate how the parameters of the health and education submodel for low-income countries are computed using the proposed approach in Section 3.6. The equations for the system are the following.

$$\text{Change in health services}: \frac{\mathrm{d}HS_t}{\mathrm{d}t} = HG_t - HD_t \tag{3.20}$$

$$\text{Change in education services}: \frac{\mathrm{d}ES_t}{\mathrm{d}t} = EG_t - ED_t \tag{3.21}$$

$$\text{Growth in health services}: HG_t = \Lambda_{\mathrm{H}r} \cdot \Delta Y_t \tag{3.22}$$

$$\text{Growth in education services}: EG_t = \Lambda_{\mathrm{E}r} \cdot \Delta Y_t \tag{3.23}$$

$$\text{Health service depreciation}: HD_t = \zeta_{\mathrm{H}r} \cdot HS_t \tag{3.24}$$

$$\text{Education service depreciation}: ED_t = \zeta_{\mathrm{C}r} \cdot ES_t \tag{3.25}$$

$$\text{Female health access}: HF_t = \chi_{\mathrm{HF1}} + \chi_{\mathrm{HF2}} \cdot \log\left(HS_t\right) \tag{3.26}$$

$$\text{Female education attainment}: EF_t = \chi_{\mathrm{EF1}} + \chi_{\mathrm{EF2}} \cdot \log\left(ES_t\right) \tag{3.27}$$

$$\begin{aligned}\text{Poor healthcare access}: HA_t^{\mathrm{POOR}} &= \chi_{\mathrm{HA1}} + \chi_{\mathrm{HA2}} \cdot \log\left(HS_t\right) \\ &+ \chi_{\mathrm{HA3}} \cdot \widehat{Y}_t^{\mathrm{POOR}}\end{aligned} \tag{3.28}$$

$$\text{Rich healthcare access }: HA_t^{\text{RICH}} = \psi_{\text{HA1}} + \psi_{\text{HA2}} \cdot \log\left(HS_t\right) + \psi_{\text{HA3}} \cdot \widehat{Y}_t^{\text{RICH}} \tag{3.29}$$

There are only two stock variables: HS_t and ES_t. The variable definition are not important for this exercise but are listed in Table C.2 in Appendix C.

We have data y_{it} for all the state variables , which are defined in the following vector with $N = 10$ elements.

$$y_{it} = \left(HS_t, ES_t, HG_t, EG_t, HD_t, ED_t, HF_t, EF_t, HA_t^{\text{POOR}}, HA_t^{\text{RICH}}\right) \tag{3.30}$$

The simulated values $\widehat{X}_{it}$ are the following.

$$\widehat{y}_{it} = \left(\widehat{HS}_t, \widehat{ES}_t, \widehat{HG}_t, \widehat{EG}_t, \widehat{HD}_t, \widehat{ED}_t, \widehat{HF}_t, \widehat{EF}_t, \widehat{HA}_t^{\text{POOR}}, \widehat{HA}_t^{\text{RICH}}\right) \tag{3.31}$$

We also assumed that each state variables has a corresponding weight.

$$w_i = (W_1, W_2, \ldots, W_{10}) \tag{3.32}$$

We have $M = 14$ parameters, which will be the decision variables that we

are optimizing to minimize the least-square error shown in Equation (3.17).

$$p_j = \{\Lambda_{\mathrm{H}r}, \Lambda_{\mathrm{E}}, \zeta_{\mathrm{H}}, \zeta_{\mathrm{C}}, \chi_{\mathrm{HF1}}, \chi_{\mathrm{HF2}}, \chi_{\mathrm{EF1}}, \chi_{\mathrm{EF2}}, \chi_{\mathrm{HA1}}, \chi_{\mathrm{HA2}}, \chi_{\mathrm{HA3}}, \psi_{\mathrm{HA1}}, \psi_{\mathrm{HA2}}, \psi_{\mathrm{HA3}}\} \tag{3.33}$$

and for each parameter, we have an upper and lower bound, $p_j^{\max}$ and $p_j^{\max}$, on the estimate.

$$p_j^{\max} = \big(\Lambda_{\mathrm{H}}^{\max}, \Lambda_{\mathrm{E}}^{\max}, \zeta_{\mathrm{H}}^{\max}, \zeta_{\mathrm{C}}^{\max}, \chi_{\mathrm{HF1}}^{\max}, \chi_{\mathrm{HF2}}^{\max}, \chi_{\mathrm{EF1}}^{\max}, \chi_{\mathrm{EF2}}^{\max}, \chi_{\mathrm{HA1}}^{\max}, \chi_{\mathrm{HA2}}^{\max}, \chi_{\mathrm{HA3}}^{\max}, \psi_{\mathrm{HA1}}^{\max}, \psi_{\mathrm{HA2}}^{\max}, \psi_{\mathrm{HA3}}^{\max}\big) \tag{3.34}$$

$$p_j^{\min} = \big(\Lambda_{\mathrm{H}}^{\min}, \Lambda_{\mathrm{E}}^{\min}, \zeta_{\mathrm{H}}^{\min}, \zeta_{\mathrm{C}}^{\min}, \chi_{\mathrm{HF1}}^{\min}, \chi_{\mathrm{HF2}}^{\min}, \chi_{\mathrm{EF1}}^{\min}, \chi_{\mathrm{EF2}}^{\min}, \chi_{\mathrm{HA1}}^{\min}, \chi_{\mathrm{HA2}}^{\min}, \chi_{\mathrm{HA3}}^{\min}, \psi_{\mathrm{HA1}}^{\min}, \psi_{\mathrm{HA2}}^{\min}, \psi_{\mathrm{HA3}}^{\min}\big) \tag{3.35}$$

So, we can formally set up the least squares problem as

$$\min_{p_j} \quad \sum_{i=1}^{10} \sum_{t=1980}^{2015} \left(\frac{y_{it} - \widehat{y}_{it}}{w_i} \right)^2 \tag{3.36}$$

$$\text{s.t.} \quad p_j^{\min} \leq p_j \leq p_j^{\max} \quad j = 1, \ldots, 14. \tag{3.37}$$

Using the island genetic algorithm, we were able to fit the health and education state variables for low-income countries. Figures 3.6.1, 3.6.1, 3.6.1, 3.6.1, 3.6.1, and 3.6.1 shows the predicted fit versus the actual fit.

In order to illustrate the actual trendlines of the state variables, we elected

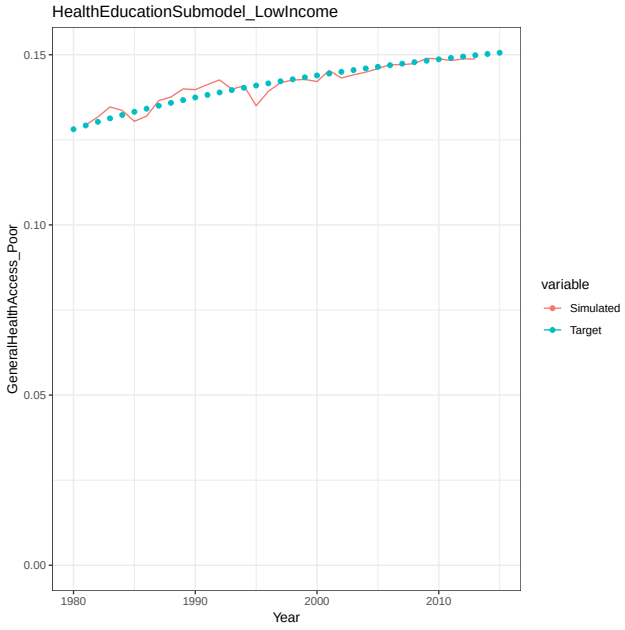


Figure 3.14: General Health Access of Poor Population in Low Income Region. Target series represent the a smoothed version of actual data points.

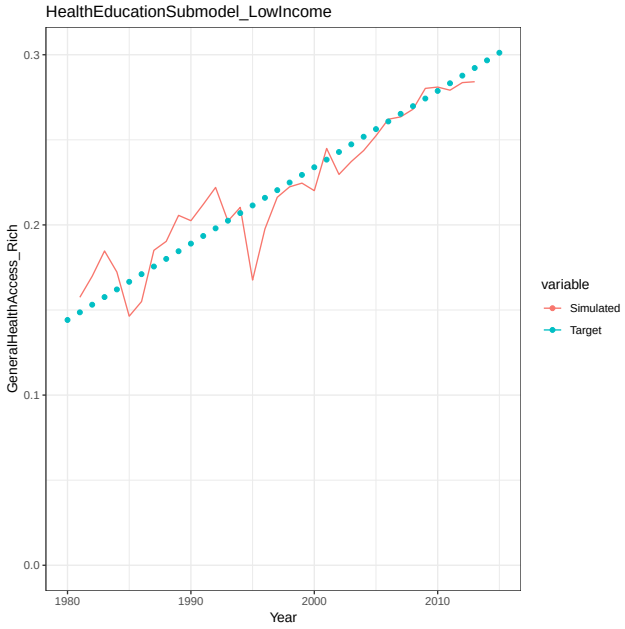


Figure 3.15: General Health Access of Rich Population in Low Income Region. Target series represent the a smoothed version of actual data points.

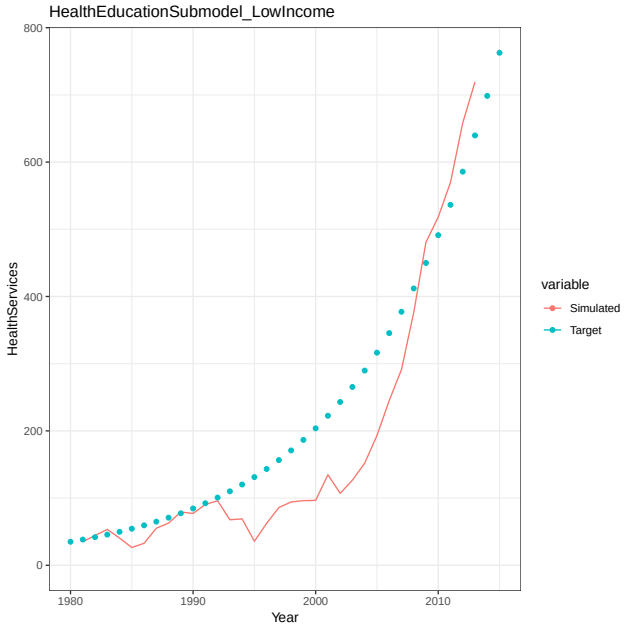


Figure 3.16: Health Services in Low Income Region. Target series represent the a smoothed version of actual data points.

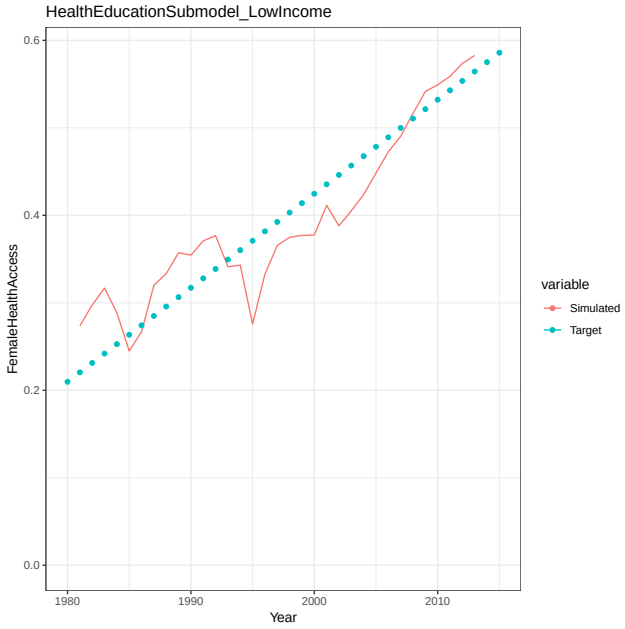


Figure 3.17: Female Health Access in Low Income Region. Target series represent the a smoothed version of actual data points.

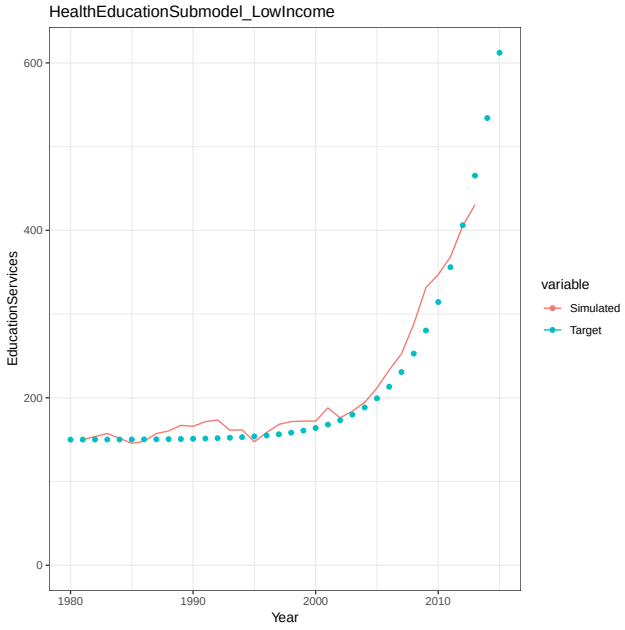


Figure 3.18: Education Services in Low Income Region. Target series represent the a smoothed version of actual data points.

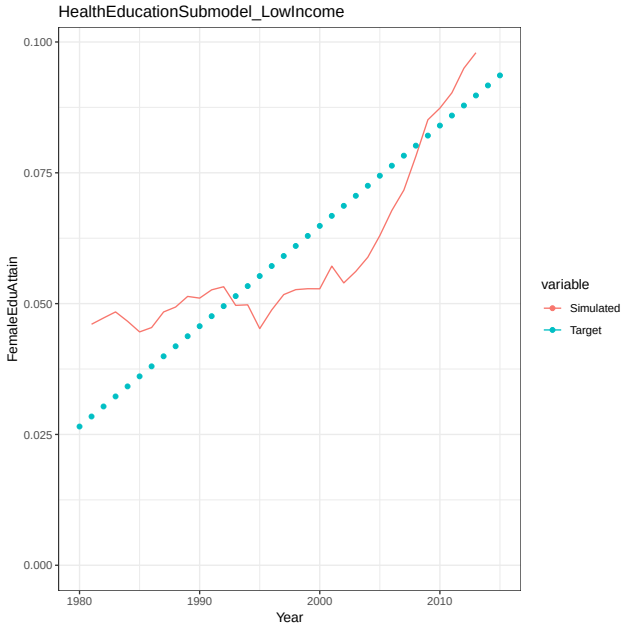


Figure 3.19: Female Education Attainment in Low Income Region. Target series represent the a smoothed version of actual data points.

to use a smoothed, denoised series instead of actual raw data of the state variables. It is also noticeable that the simulated results are very noisy. We did not smooth out the exogenous input data of the health and education submodel. The noise in the simulated values corresponds to the actual economic output and income per capita data, which are defined as aggregate Gross Domestic Product (GDP) and GDP per capita of all the countries in the low-income region, respectively. We also calibrated the health and education submodel for middle-income and high-income regions (see Appendix D).

3.7 Conclusion

The main contribution of this work is to present an endogenized population submodel that is fully integrated into a sustainability and health model. Other contributions include the ability to model disparities of food and healthcare access between socioeconomic groups and income regions. Since resource scarcity typically affects the poorer populations more, it is important to include these features into the equations of motions that govern global population dynamics.

The results from the submodel calibration show that it is possible to calibrate all submodels that were presented. Once the parameters of each submodel are all estimated, we can proceed to integrate the submodels into the overall model with endogenous feedback. To achieve full integration, we propose a

methodology that would make it possible to interface all the submodels and observe the multiple feedback loops that control the behavior of the system. The second step would require us to re-estimate all the parameters in the whole model using the submodel parameter estimation as an initial guess.

Chapter 4

Integrated Markovian Modeling of Climate-driven Migration and Urbanization in the United States

The motivation for this chapter is to demonstrate the use of Markov Chains for origin and destination data of human mobility. Furthermore, we show how we integrated the Markov Chain model with a system dynamics model that features a feedback loop that encompasses the global climate and population dynamics of the United States. The chapter first presents the background, then lays out the model formulation. Afterward, we show results from the model

based on parameterization from empirical data.

4.1 Background

Urbanization is a phenomenon that is occurring all over the world as more people are finding better economic, social, and housing opportunities in cities [93]. The attracting forces of migration are commonly known as *pull* factors. Besides the pull factors that attract people to cities, there are also *push* factors, such as the environmental changes and economic uncertainty, that force people to leave their existing residence to pursue stability for themselves and their families. However, given the exacerbation of global climate change in recent times, environmental push factors will become a greater influence on mobility in the United States as sea-level rise displace residents in low altitude regions near the coast and impacts on agricultural production and rural economy motivate residents to pursue opportunities in cities [94].

Given sea-level rise (SLR) projections for the next hundred years range from 0.3 meters to 2.0 meters, the United States must consider the possibility of permanent migration of coastal inhabitants and prepare for adaptation to climate change. Depending on the magnitude of SLR, it has been estimated that 4.2 million to 13.1 million people could be at risk of inundation [95]. Some have already examined the internal migration flows caused by climate

change that may occur if the US does not take steps in reinforcing its current infrastructure to protect the affected regions [96]. It is important to model these relationships in order to quantify the cost and damages that will be incurred.

With increased migration flows into inland areas, existing cities will face an additional burden on their infrastructure and social services. Furthermore, increased urbanization also has environmental implications, such as faster urban development that contributes to global warming (i.e. heat island effect). Faster urbanization may also accelerate industrialization that contributes to greenhouse gas emissions. However, other studies have shown that urbanization may lead to lower emission per capita of greenhouse gases (GHG) [97]. The United States accounts for 15% of all CO_2 emissions in 2014 [98]. That contribution to emissions will drive the pace of global warming. Hence, urbanization is a systemic issue that contains reinforcing feedback loops that couples population dynamics and climate change.

Other other frameworks have explored the feedback between migration and climate change by looking at the epidemiological implications of climate change impacts on infectious diseases and migration [99]. System dynamics set up an appropriate mathematical structure to model migration [100].

We explore the internal migration patterns of the United States based on tax exemption data from the United States Internal Revenue Service (IRS)

and demographic data from the Census Bureau. The study is carried out using as a system dynamics model that embeds a Markov Chain to determine the destination probability of internal migration.

4.2 Data

Internal migration flows were extracted from the US Internal Revenue Service's Statistics of Income (SOI), and county population size was extracted from the United States Census Bureau. We are using the 2013 Urban-Rural Classification Scheme from the National Center of Health of Statistics (NCHS) to designate the extent to which a county is urban or rural. We assumed counties are urban if they contain a large metropolitan statistical area (MSA) with at least 1 million people or more, or if they are a fringe county of MSA with 1 million or more people. The coastal areas were designated based on their geographic proximity to the Atlantic Ocean, Pacific Ocean, and the Gulf of Mexico. The list of coastal counties was obtained also obtained from the US Census Bureau.

Given these assumptions, we were able to aggregate the populations of all counties in the US into 4 different groups:

- **Coastal, Urban** – This includes urban areas that border the Atlantic Ocean, Pacific Ocean, or the Gulf of Mexico.
- **Non-coastal, Urban** – Urban areas that are not considered coastal.

- **Coastal, Rural** – Rural areas along aforementioned bodies of water.
- **Non-coastal, Rural** – Rural areas that are not considered coastal.

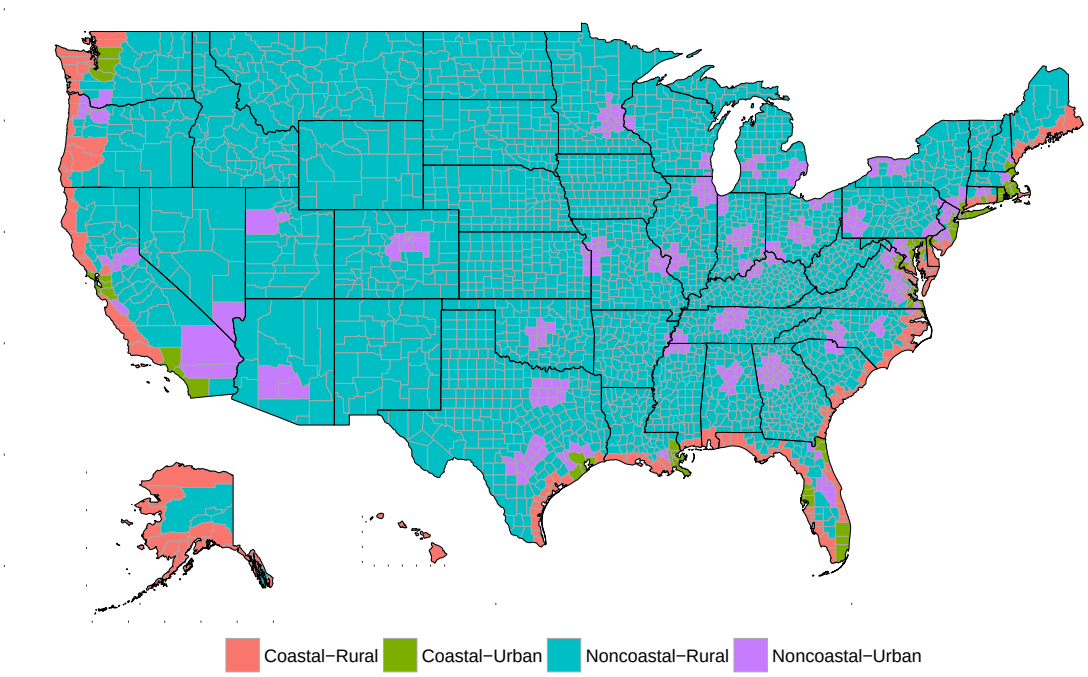


Figure 4.1: We classified each county into four groups which are shown on the map.

The initial parameters of the model such as population size and migration flows can be determined using the data from the NCHS and US Census Bureau. We used population estimates from the US Census Bureau to calculate the population change rate of each aggregate region. Specifically, we collected data on county-level birth, death, and international migration rates and used that information to calculate and parameterize the growth rate of each population group. All data and projections were processed in the R statistical environment.

Migration patterns were extracted from the Statistics of Income (SOI) dataset created by the Internal Revenue Service (IRS). This dataset includes year-to-year address changes from individual income tax statements. Since the majority of residents have to file their income taxes, internal migration patterns are reasonably represented in the SOI dataset. Nevertheless, there are limitations in the SOI dataset; seasonal and return migration are also captured which makes it difficult to distinguish from long-term and forced migration that is caused by economic and environmental events.

Migration flow for 2015-2016 is shown in Figure 4.2. We can see that the majority of internal migration takes place within the aggregate regions which probably reflects a short-distance move. Migration within aggregate regions also includes inter-county migration. We are mostly interested in the rural-urban migration flows. The distribution of year-to-year migration is fairly consistent from 2011-2016. The migration flows for 2015-2016 can be better represented in the chord diagram in Fig. 4.3 where the width of links represents the volume size of flow.

4.3 Motivation and Background

We developed a computational model that focuses on the theoretical relationship between migration flow and climate-driven events. We want to know how many

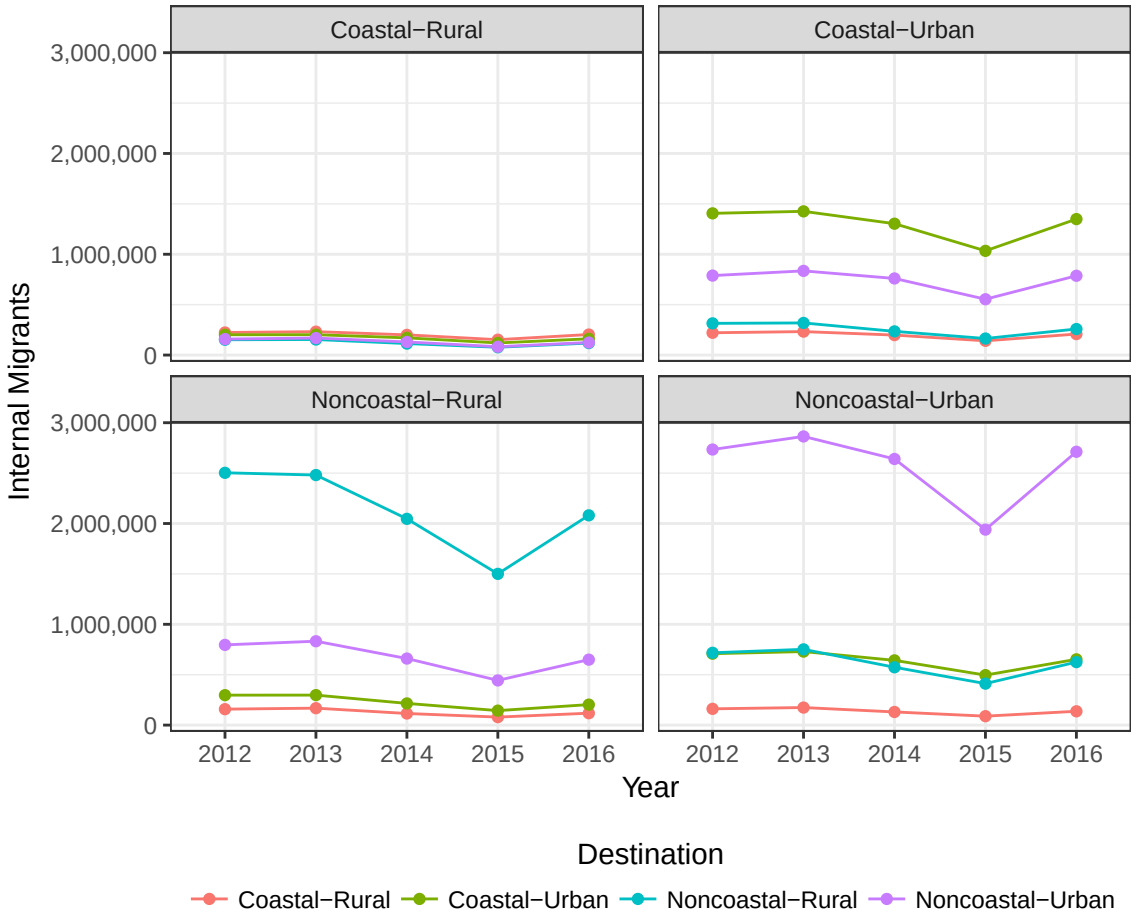


Figure 4.2: The four plots show the 2012-2016 internal mobility (county-to-county flows) in the United States for our aggregate groups. For clarification, the top right plot labeled "Coastal-Urban" represent the migration from coastal, urban counties, and the four destinations are color coded. In that particular plot, the highest migration flow is within the coastal-urban aggregate region (i.e., coastal-urban to coastal-urban) which is represented by the green line. This is based on the IRS Statistics of Income (SOI) dataset.

coastal residents will move after their house is displaced by sea-level rise. In rural areas, we want to know how many residents' livelihoods are threatened due to climate change impacts on agricultural production, and the impact that will have on rural-to-urban migration. The combined effect of these two events

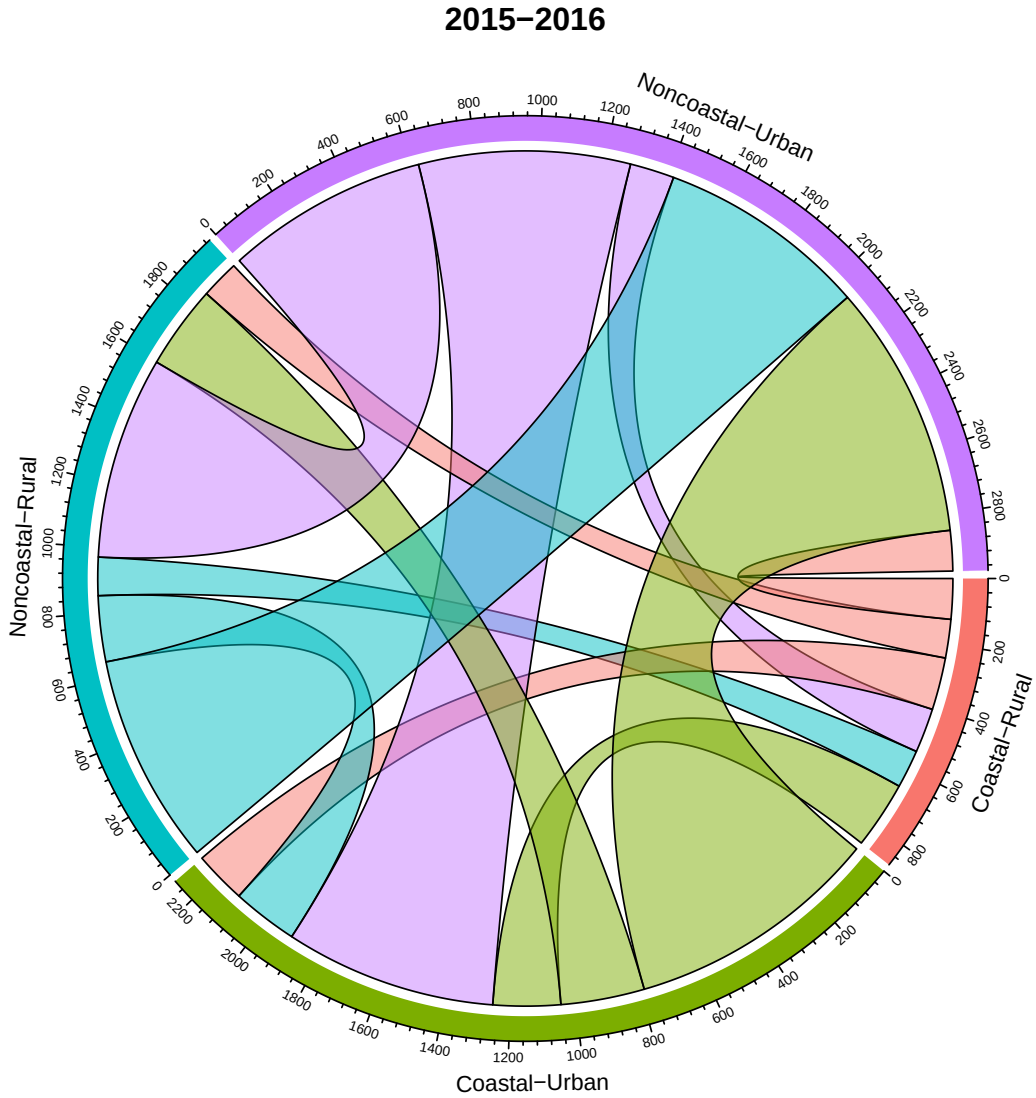


Figure 4.3: The inter-regional migration flows are shown here where the widths are scaled based on migration size. The tick marks represent 1000 people. We do not include migration within regions, which is the majority of internal migration in the US.

will impact urbanization of inland cities.

The implementation includes a Markov Chain model with integrated feedback loops as a proof-of-concept that capture the dynamics that drive climate change

and urbanization. The structure of our model is shown in Figure 4.4. The four groups that were mentioned in the previous section are utilized in our model. Based on the data collected, we were able to calculate the initial conditions and parameters associated with population and migration dynamics.

We included two important feedback loops that connect climate change and urbanization. The first feedback loop includes migration from coastal to non-coastal areas due to inundation from sea-level rise. We've also included a second feedback loop that connects agricultural production that is sensitive to climate conditions with migration flows from rural-to-urban regions. Based on the IRS SOI data, we were able to delineate future migration trends based on current migration patterns. However, in a more extreme climate change context, we can argue that the two feedback loops included in our model will magnify the aforementioned population movements towards safer regions, i.e., non-coastal urban regions.

The computational model is developed based on a Markov Chain with feedback that is commonly used in discrete-time simulations of the migration system. The system of equations that govern the dynamics of our computational model is shown in the next section. The model is implemented in the R environment as a system of difference equations.

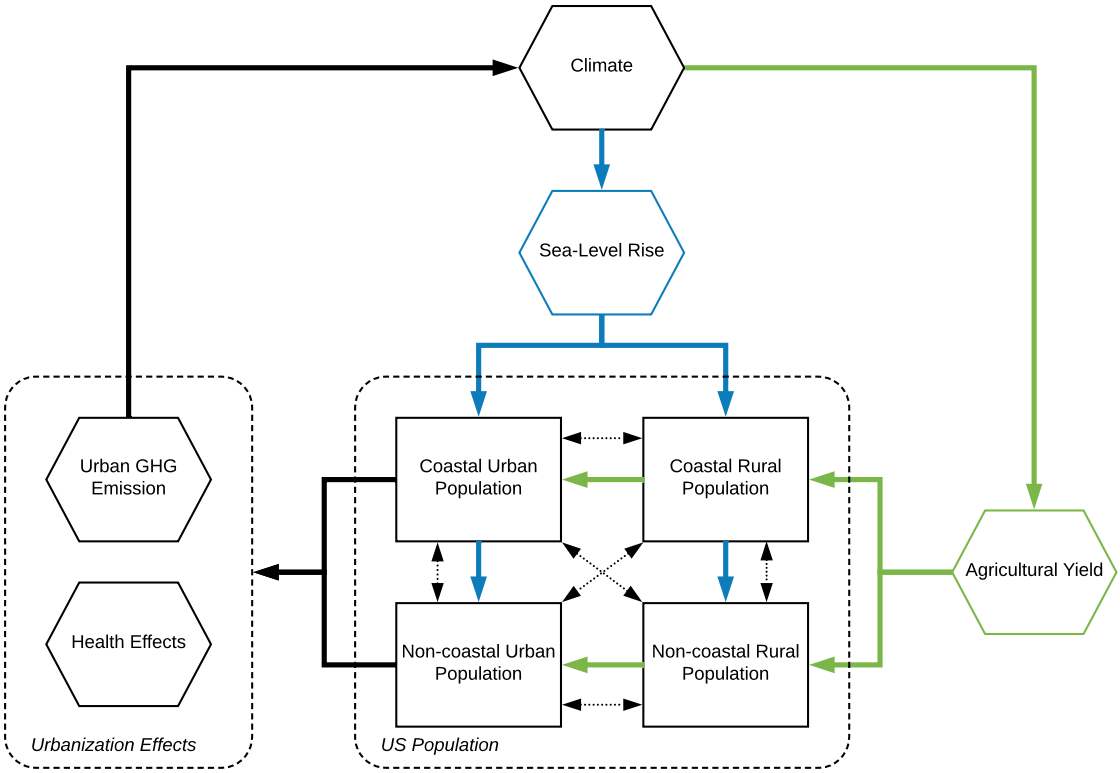


Figure 4.4: The conceptual causal links of our computational model is shown along with the two primary feedback loops. Urbanization and its impact on GHG emissions will impact climate conditions. Subsequently, climate-sensitive events such as sea-level rise (blue arrows) and agricultural yield (green arrows) will push residents in these vulnerable regions to move to a more stable region in the US. The dotted arrows represent typical migration patterns that occur in a "business-as-usual" situation.

4.4 Markov Chain Integration in a System Dynamics Model

We define the population of the four region i where $i \in \{\text{Coastal-Rural} = 1, \text{Coastal-Urban} = 2, \text{Noncoastal-Rural} = 3, \text{Noncoastal-Rural} = 4\}$ at time

$t+1$ as the following

$$P_i(t+1) = P_i(t) + B_i(t) - D_i(t) + \sum_{j \neq i} M_{ji}(t) + G_i(t) \tag{4.1}$$

where P_i is the population, B_i is the birth count, D_i is the death count, M_{ji} is the migration flow from region j to region i within the US, and G_i represent net migration from other countries into the US. We can also represent variables B_i, D_i, M_{ji}, and G_i as rates instead of count variables. The birth, death, internal migration, and international migration rates can respectively be represented by b_i, d_i, m_{ji} and g_i. Therefore, the population of each region can be recast as

$$\begin{aligned} P_i(t+1) &= P_i(t)(b_i(t) - d_i(t) + g_i(t)) + \sum_{j \neq i} P_j(t) m_{ji}(t) \\ &= \phi_i(t) P_i(t) + \sum_{j \neq i} P_j(t) m_{ji}(t) \end{aligned} \tag{4.2}$$

The variable ϕ_i represent the growth rate of each region minus the internal migration. We used a linear model to forecast the population growth rate ϕ_i for each region.

Given the dynamics of Equation (4.2), we can model the population size of each region as a Markov Chain Model (MCM). The exogenous inputs, i.e., the growth rate is defined as a diagonal vector $\mathbf{\Phi}(t) = \text{diag}(\phi_1(t), \ldots, \phi_4(t))$ where

the elements correspond with the rate of natural increase plus international net migration for each region. Based on the birth, death, and immigration rates that were calculated from the US Census Bureau estimates from 2011-2016, we were able to linearly project the population change rate with respect to time for all 4 regions. Hence, births, deaths and international migration are considered to be exogenous in our model. The state (population) of the system at time t is stored in the vector $\mathbf{p}(t) = [P_1(t), \ldots, P_4(t)]^T$. Our population can be modeled as

$$\mathbf{p}(t+1) = \mathbf{p}(t)\mathbf{\Phi}(t) + \mathbf{p}(t)\mathbf{M}(t). \tag{4.3}$$

The transition matrix M is the collection of transition probabilities for internal migration between two regions, and the dynamics of the transition probabilities is of great interest to us. Internal migration by definition is a closed system where a resident's decision to stay within the same region or move to one of the other 3 regions are mutually exclusive and exhaustive events. In other words, $\sum_j \mathbf{Pr}(X_{t+1} = \text{live in region } j | X_t = \text{live in region } i) = 1$. Therefore, the internal migration transition matrix M must be a stochastic matrix where all the elements in each row sum up to 1. We can maintain the stochastic properties while altering the transition probabilities with the following formulation,

$$\mathbf{M}(t) = \mathbf{M}_0 + \mathbf{M}_S(t)\Delta_S(t) + \mathbf{M}_A(t)\Gamma(t). \tag{4.4}$$

The scaler coefficients Δ_S and Γ represent the sea-level and agricultural production impacts on migration, respectively. M_0 is the initial transition matrix that is parameterized with data from the IRS SOI dataset for migration in 2011-2012. We calculate the initial transition probability of internal migration between all pairwise regions. We can explicitly define the initial transition matrix as

$$\mathbf{M}_0 = \begin{bmatrix} \frac{M_{11}}{P_1} & \frac{M_{12}}{P_1} & \frac{M_{13}}{P_1} & \frac{M_{14}}{P_1} \\ \frac{M_{21}}{P_2} & \frac{M_{22}}{P_2} & \frac{M_{23}}{P_2} & \frac{M_{24}}{P_2} \\ \frac{M_{31}}{P_3} & \frac{M_{32}}{P_3} & \frac{M_{33}}{P_3} & \frac{M_{34}}{P_3} \\ \frac{M_{41}}{P_4} & \frac{M_{42}}{P_4} & \frac{M_{43}}{P_4} & \frac{M_{44}}{P_4} \end{bmatrix} \tag{4.5}$$

where the elements $\frac{M_{ij}}{P_i}$ are calculated based on the proportion of the population in region i migrating to region j.

The matrix M_S adjusts the initial transition probabilities of coastal regions in M_0 based on the sea-level while maintaining the stochastic properties of M. Specifically, this matrix adjusts the migration from the non-coastal to coastal regions. We weighted the matrix M_S so non-coastal, urban areas receive more

migration. This matrix is defined as the following.

$$\mathbf{M}_S = \begin{bmatrix} -1.5 & -1.5 & 1 & 2 \\ -1.5 & -1.5 & 1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.6}$$

The matrix M_A adjusts the transition probabilities of M_0 with respect to agricultural output. Specifically, M_A represents the change in the migration probability from rural to urban regions due to the adverse impacts of climate change on agriculture-based economies in rural counties.

$$\mathbf{M}_A = \begin{bmatrix} -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.7}$$

The element values of M_S and M_A are based on theoretical effects of sea-level displacement and agricultural production. These are assumed values that meant represent the effect-size on the transition probabilities of internal migration by sea-level rise and agricultural production.

The scaler coefficients that adjust the magnitude of migration due to sea-level rise and agriculture in Equation (4.4), M_S and M_A, are calculated as

$$\Delta_S(t) = \delta S(t) \tag{4.8}$$

$$\Gamma(t) = \frac{\gamma}{A(t)} \tag{4.9}$$

The coefficients α and β are parameters that will have to be adjusted during our scenario runs to reflect different impact magnitudes of sea-level S and agricultural production A. The equations that govern the sea-level and agricultural production trends are

$$S(t+1) = S(t) + \alpha T(t) \tag{4.10}$$

$$A(t+1) = A(t) - \beta T(t) \tag{4.11}$$

where the coefficients α and β represent the change rate of sea-level rise and agricultural production with respect to mean temperature (i.e., climate). We are assuming that agricultural production is inversely correlated to climate. Studies have suggested that climate change has caused more incidents of extreme weather events that may cause droughts [101].

We also based our temperature model, on a simplified set of equations that model relationship between CO_2 emission and climate. The CO_2 that is emitted

will cause an increase in radiative forcing F_{CO2} of CO_2.

The equation for temperature can be represented as

$$T(t+1) = T(t) + \Delta_T(t) \tag{4.12}$$

where the first-order temperature change Δ_T is defined as

$$\Delta_T(t) = \lambda F_{\mathrm{CO2}}(t) + F_{\mathrm{OTHER}}(t). \tag{4.13}$$

We also include radiative forcing from other greenhouse gases (GHG) as F_{OTHER}. We model the CO_2 radiative forcing F_{CO2} as a function of total carbon dioxide emission:

$$F_{\mathrm{CO2}} = 5.35 \ \ln\left(\frac{C(t)}{C_0}\right) \tag{4.14}$$

Finally, the total CO_2 is simply the sum of the emission rate from all four regions. The constant C_0 represents the reference (initial) carbon dioxide concentration. We assume the emission rate per capita $\widehat{C}_i$ varies between the regions and is exogenous. The variable EC represent the CO_2 emissions from other countries.

$$C(t) = \sum_{i=1}^{4} \widehat{C}_i P_i(t) + EC(t) \tag{4.15}$$

4.5 Migration Results

The model presented here is simply a framework for an integrated model. We ran the model with the population data collected from the IRS and US Census Bureau to simulate the potential trajectory. As mentioned before, the birth, death, and international migration rates are considered exogenous, which led to a similar projection with the US Census Bureau (see Figure 4.5). We parameterized a linear growth model based on birth, death, and international immigration rates of each region listed in the data from the US Census Bureau.

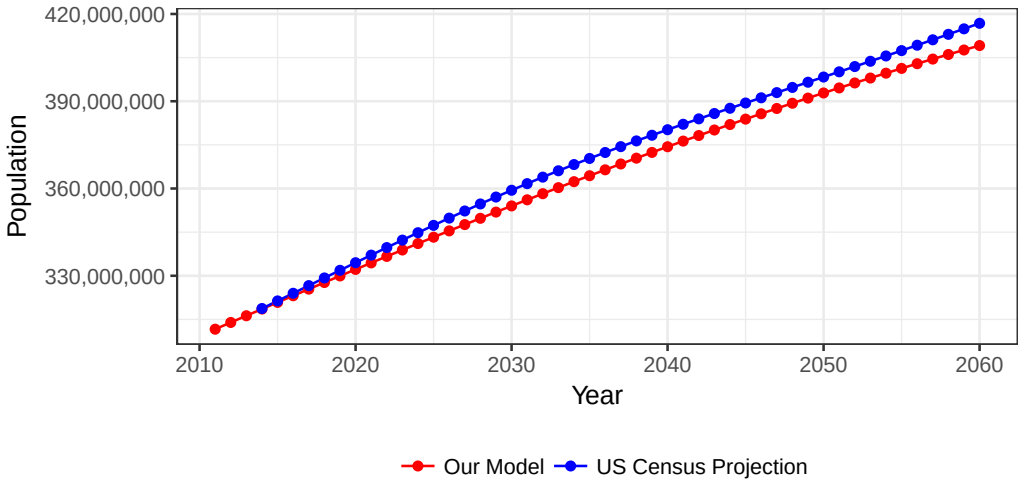


Figure 4.5: Comparison of our working model and the United States Census Bureau projections up to year 2060.

The population trends from 2011-2016 for each region are shown in Figure 4.6, and this captures the hypothesized relationship between climate change and migration. We can see that the non-coastal, urban region is growing faster than any other region. This is consistent with our assumptions that people will

migrate towards cities that are inland.

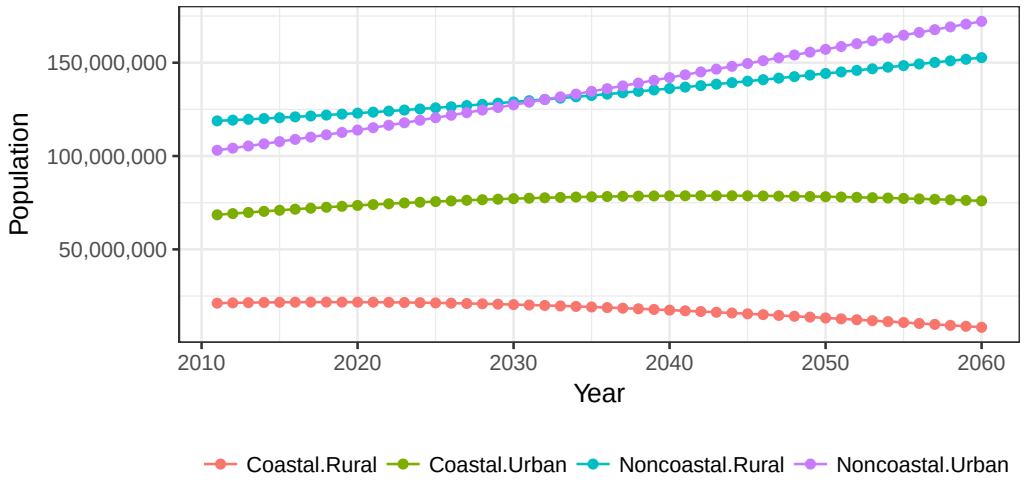


Figure 4.6: The projections of our working model for each group up to the year 2060.

The results are

4.6 Conclusion

The results presented in Section 4.5 demonstrate one possible trajectory for population dynamics. From the results, we see that it is possible to capture long-term trends using a dynamic model that have other factors with relevant feedback. Since climate change effects are a nonstationary process, we must rely on careful calibration of feedback models to understand effects on health systems, population mobility and growth, and greenhouse gas emission rate increase. We used theoretical relationships of temperature with agricultural

production and sea-level rise to estimate a possible effect on internal migration.

The main contribution of this model is an integrated system dynamics model that embeds a Markovian process. We do not consider pull-factors, such as economic opportunity, which may alter future migration patterns, but this may be difficult to include in our model since this requires additional assumptions based on individual decision-making. One may integrate individual-level, agent-based decision functions to this model, but this will inevitably increase the computational requirements. We introduce a Markov Process which is based on empirical probabilities.

At the moment, the model only includes the aggregate population dynamics. The spatial scale could be refined later if we decide to increase the granularity of the model to include county-level systems. At the current spatial scale of the model, it would be difficult to model infrastructure, social, and economic systems that might influence internal migration. We can increase the granularity of the model to include county-level dynamics. This would significantly increase the dimensionality of the problem. The model can be refined by including the variability of sea-level rise for each coastal region. Also, we can extend the model to include geographic differences between each coastline community using geographical information systems (GIS).

Chapter 5

Conclusion and Outlook

The work that was presented in Chapters 2, 3, and 4 are just a few examples of how system modeling could be integrated to provide a richer understanding of how the system works as the sum of its parts. The need for integrated climate models has long been discussed due to the complexity of the interconnectivity between the natural environment and human systems. Although it is commonly understood that "all models are wrong," it is still difficult to understand the true system without attempting to model the interactions of the subsystems that make up the entire system.

Furthermore, we see that sustainability have a direct impact on public health since these issues are far from being mutually exclusive topics. In a world that has finite resources, nutritional intake of poorer populations will be restricted and put the health of millions in jeopardy as resources become

scarce. Social consequences of growing inequality will also impact poorer populations and their access to basic healthcare services. This is especially true in low-income countries that have yet or are in the midst of transitioning economically into middle-income countries. We show in Chapter3 that an integrated model will allow researchers to understand the effects on population growth from a systems perspective due to the endogeneity of population submodel.

Even in wealthier countries, such as the United States, we see that the effects of climate change will drive existing populations in coastal communities and rural communities towards cities, causing faster urbanization. As increased urbanization occurs, we will face increased demand on existing urban systems and, to a lesser extent, crowding. In Chapter 4, we, again, see that integrated modeling allow us to endogenize migration dynamics and understand feedbacks between climate and human mobility.

The primary contribution of this thesis is to show the potential for integrated models to capture the behavior of complex systems. By endogenizing certain facets of the system, such as population dynamics, we are able to observe the systemic issues that affect the observed components. This can be extended universally to any other issues. We also show that complexity, such as the collaboration behavior of clinical trials, can be captured with the necessary data. This complexity can then be related to successful outcomes like drug approvals.

5.1 Future of System Dynamics in the Age of Big Data

Given the incredible explosion of data-availability, system dynamics modelers are now in a unique position that was previously impossible to explore many social and health issues with the aid of evidence and data. For much of the history in system dynamics, modelers had solely to rely on anecdotal and intuitive logic to formulate the causal relationships between variables because there did not exist the data to support the intermediate linkages that connect the input and output variables. With advancements in data collection technology, researchers are now able to incorporate empirical trends into the model-building exercise.

The models presented in this thesis demonstrate the potential for integrating a multi-component model with data from different sources. System dynamics is a useful medium to tie disparate sources of data together to discover causal relationships between factors that are usually measured independently. The aid of machine learning also helps us identify the prominent factors that drive the system. However, we must supplement these analyses with integrated system dynamics models to truly understand the underlying causal dynamics.

The conclusions in Chapter 2 demonstrate the need for interdisciplinary collaboration and diversity of ideas is paramount to accomplishing successful

research in medicine. This idea can be extended to systems research in public health and sustainability. For centuries, science has been exclusively conducted at academic institutions by siloed disciplines that safeguard their research by constraining the flow of knowledge. However, as we traverse into this brave new world, where data, information, and knowledge are freely available – we must learn how to maximize the utility of this unbounded, limitless resource. System thinking and modeling provides a medium for converting data into information which is then synthesized into knowledge.

5.2 Summary of Future Work

Since most of our research is an ongoing collaborative process, we will outline the next steps for each project. The progress and results presented in this thesis encompass the interdisciplinary work with other researchers.

5.2.1 Clinical Trials and the System Methodology

In Chapter 2, the network analysis was the synthesis of a culmination of work with the MIT Collaborative Initiatives using the system methodology. Prof. Sauleh Siddiqui and Prof. Takeru Igusa are promoting many of the ideals of the MIT Collaborative Initiatives by engaging young and mid-career academic leaders throughout Johns Hopkins in using systems approaches for grand challenge

problems. Some of these projects and their leaders, which have all taken on a broader multidisciplinary systems approach during the past 18 months, include work on: Native American youth (Allison Barlow, Teresa Brockie), safe and equitable testing and deployment of autonomous vehicles (Johnathon Ehsani), suicide in the U.S. (Emily Haroz, Elizabeth Stuart), violence against children in low- and middle-income countries (Paul Bolton), disparities in the U.S. health system (John Jackson), and promoting mental and physical activities to enhance healthy aging (Atif Adam, Michelle Carlson).

5.2.2 Multi-component World Population and Sustaina Model

In Chapter 3, we propose an integrative method for modeling population dynamics. In order to aid the transition of developing countries that are undergoing economic change, the Bill and Melinda Gates Institute for Population and Reproductive Health will continue to employ models in exploring large-scale population dynamics in the context of climate change and economic inequity.

5.2.3 US Migration Model

In Chapter 4, we develop a conceptual model with feedback loops. Along with collaborators at the Bloomberg American Health Initiatives, we will continue

work on modeling migration patterns and the effects on public health and urbanization.

Appendix A

Coupling Existing Models via Intermediate Temperature Model to Study Repeated Hazards

In this appendix, we will introduce a framework for an integrated model and develop an intermediate model that couples two models of different domains and scales in order to investigate systemic effects of repeated heatwaves on an urban environment. The *intermediate model* is a distributed-lag regression model that connects a climate model with an agent-based model (ABM) of city residents. Based on the outcome of the model, we can propose a set of interventions to aid in disaster preparedness and infrastructure resilience. Specifically, we want to understand what type of market or policy interventions

will help reduce the risk of excess mortality rates associated with heatwave events. This could model could answer questions such as, “How would subsidizing window-unit air conditioners impact reduce excess mortality?”

The research in this thesis stems from a necessity to integrate an agent-based social network model and an atmospheric climate model that predicts temperature. We examined the mid-Atlantic region of the United States. For the past decades, there has been a focus on developing models that endogenize physical and social processes related to climate and environmental change [102]. In order to harness the modeling progress of traditional fields such as meteorology and sociology, one must be able to couple two separate and disparate models using an interfacing model that that is logically and causally consistent with the application domain. In our case, we will focus on this appendix is to conduct a time series analysis between indoor and outdoor temperatures since this links the heat effects on human health. To aid with this analysis, we utilized a distributed lagged regression model that predicts indoor temperature based on outdoor temperatures.

A.1 Background

Integrated models of resiliency due to hazardous events have been widely discussed [103]. Other studies have investigated the community resilience

from an integrated systems perspective [104].

As climate change increases the frequency and intensity of heatwaves [105, 106], cities will inevitably need to take measures to defend against rising mortalities associated with heatwave events. The occurrence of heatwaves varies geographically but will affect certain regions that were already prone to heatwaves, such as the mid-Atlantic region. Heatwaves are generally defined as a period of relatively extreme, hot temperatures. Relative humidity also plays a large factor in the health risks of heatwaves. High temperatures are associated with People spend the majority of their time indoors. Indoor temperatures also vary depending on building type, area, and behavior of residents [107].

This research was conducted using an interdisciplinary approach to develop a novel integrated model that includes utilizing an agent-based model to simulate the learning behavior of residents and their willingness to seek air-conditioned refuge based on their social network. The data used in this study is based in Baltimore, Maryland, USA and their respective social connections (e.g. family, friends). Similar studies have been done in other cities [108], but none has been done for Baltimore. Additionally, this study will also integrate a climate model that predicts the outdoor temperature which will serve as an input to our indoor-outdoor temperature model.

The contribution of our indoor-outdoor temperature model in this thesis is that it serves as an intermediate model to interface between the ABM and

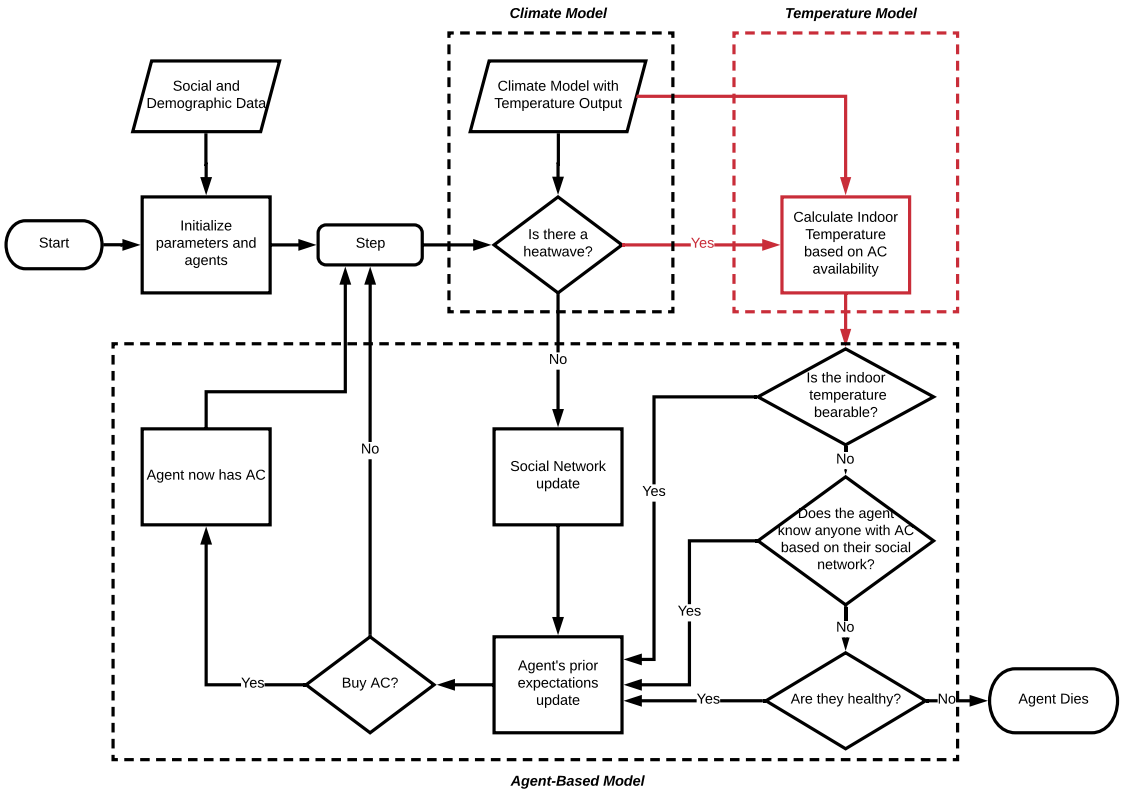


Figure A.1: Conceptual framework of the integrated heatwave model. The integrated model that is shown here requires an intermediate indoor-outdoor temperature model that bridges the climate-model and ABM (submodel is highlighted in red). The climate model feeds into the agent based via our intermediate model (one-way interaction).

climate model.

The final objectives of the project are to understand

1. the complex interplays between preparedness and resilience of a community while minimizing the near misses and false alarms,
2. influences of resource constraints and market failures on mitigation decision-making,
3. interactions between infrastructure and building hardening, land-use change, and regional resilience,
4. regional climate adaptation decision on regional resilience, and finally,
5. integration of a multidisciplinary approach to regional resilience.

This appendix will focus on objective five. This appendix proposes to couple two multiscale, dynamic models using a statistical time-series model.

A.2 Data

The indoor temperature profiles were based on the data from [109] which were collected from homes in Baltimore that have no air conditioning. We analyzed one representative home with no air conditioning with hourly measurements for 151 hours (i.e. = $N = 151$).

Series	Mean	Median	Minimum	Maximum
Indoor Temperature	34.27	34.16	32.76	35.90
Outdoor Temperature	26.92	26.81	22.18	33.86

Table A.1: Summary statistics of indoor and outdoor temperatures in Celsius.

The demeaned series (i.e., $X_t - \overline{X}$ where $\overline{X}$ is the mean of the time series) for a house with no air conditioning is shown in A.2.

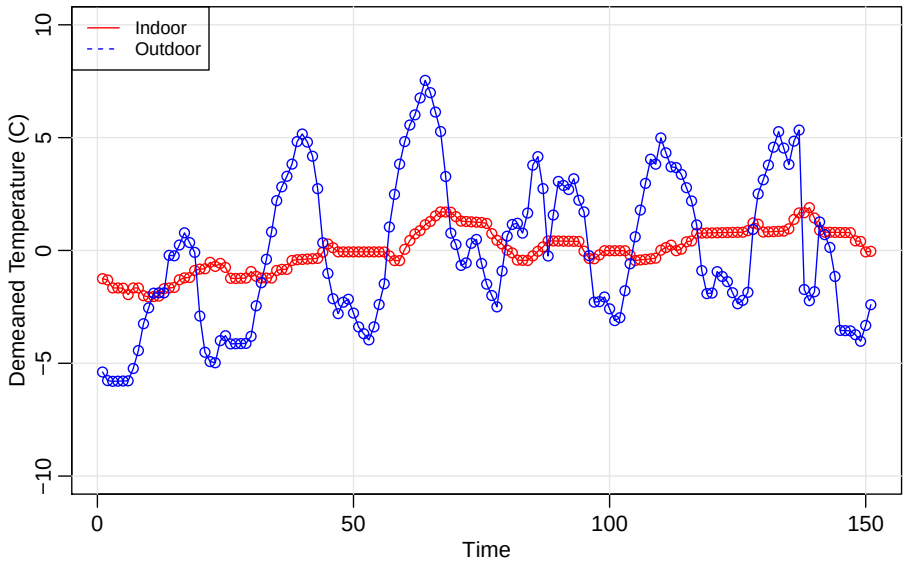


Figure A.2: Demeaned temperature profiles

From Figure A.2, we notice that there is a strong cyclical period of 24 hours with outdoor temperature, which reflects the diurnal cycle. The series is also not quite stationary since we see the outdoor increase up to hour 100 and decrease afterward while the indoor temperature series continue increasing throughout the time domain. This cycle can be broken down into daily profiles (shown in Figure A.3). Here we can see a clearer trend for the cyclical behavior.

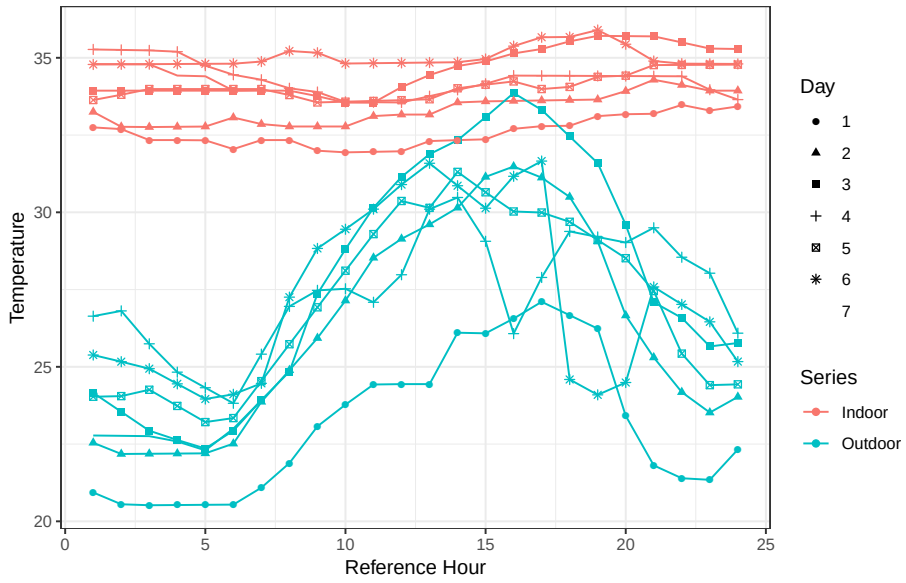


Figure A.3: Daily temperature profiles

From the time series data we were able to observe a lagged correlation within the data. We were able to calculate the cross-correlation between indoor temperature and the lagged outdoor temperature time series using the well known Pearson correlation function,

$$\text{corr}\,(X_{t-k}, Y_t) = \frac{\sum_{i=1}^{n}(X_{i(t-k)} - \overline{X})(Y_{it} - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_{i(t-k)} - \overline{X})^2 \sum_{i=1}^{n}(Y_{it} - \overline{Y})^2}}. \tag{A.1}$$

Figure A.4 shows the cross-correlation function for different time lags k. Figure A.5 explicitly shows the scatterplot the indoor and outdoor time series with respect to different lag sizes. As we can see from both figures, the peak correlations are at a lag of six hours. In other words, the indoor temperature at time t has the highest correlation value with the earlier outdoor temperature

at $t - k$.

This is consistent with the Diurnal temperature variation between indoor and outdoor temperatures. As a result, we would need to select a time-series model with delayed covariates (outdoor temperature) to determine time-specific indoor temperatures.

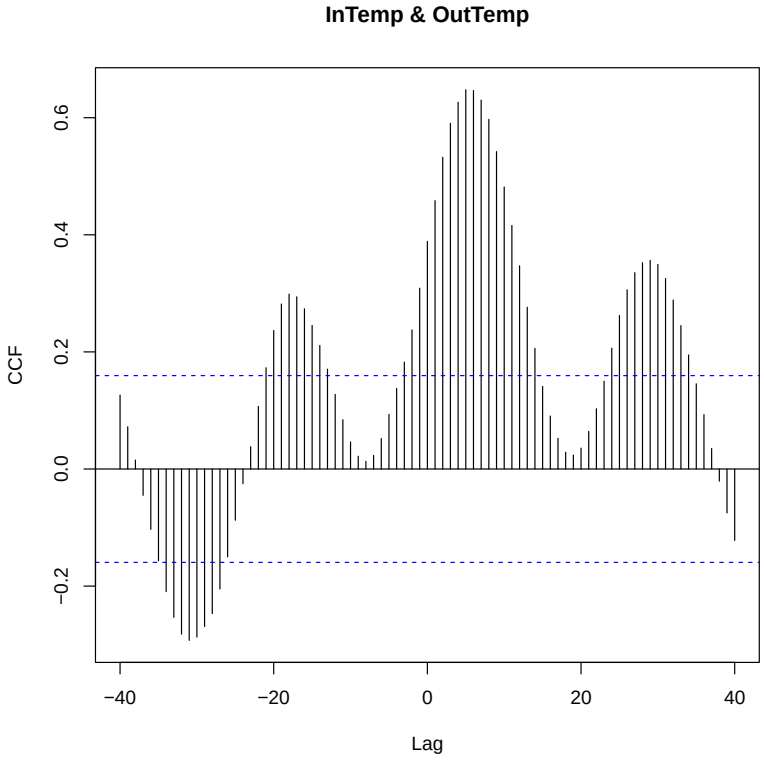


Figure A.4: Cross-correlation plot of indoor and outdoor temperatures. The lag (k) represents the correlation between an earlier outdoor temperature X at $t - k$ and indoor temperature Y at t. We see that the highest cross-correlation between lagged indoor and outdoor temperature is six hours.

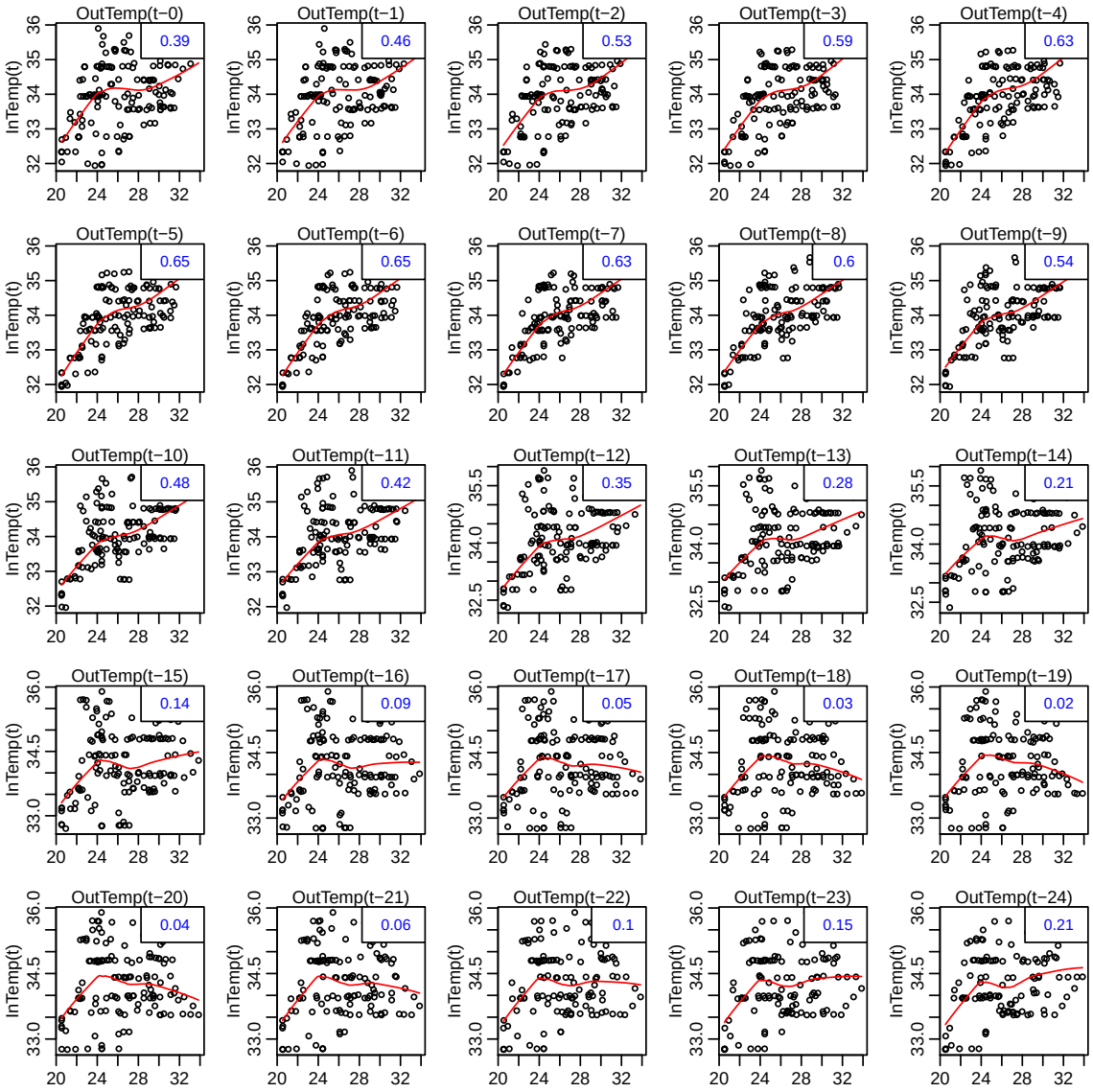


Figure A.5: Scatterplot of Indoor and Outdoor Temperature.

A.3 Developing the Indoor-Outdoor Tempera Model

We were able to develop a four demeaned, distributed lagged regression models with varying orders of lag. The functional forms shown in equations (A.2),(A.3),(A.4), and (A.5). Selecting this model was based on an intuitive understanding of the delay between indoor and outdoor temperature. We used a linear structure for this model since it would allow for faster parameter estimation. The model also assumes that there are no other time-dependent factors that affect the indoor temperature estimate, which for the integrated model is appropriate since we are assuming the agent's houses are consistent in structure, ventilation, and insulation.

$$Y_t = \alpha + \beta_1 \left(X_t - \overline{X}\right) + \beta_2 \left(X_{t-6} - \overline{X}\right) + \beta_3 \left(X_{t-12} - \overline{X}\right) + \beta_4 \left(X_{t-24} - \overline{X}\right) \tag{A.2}$$

$$Y_t = \alpha + \beta_1 \left(X_t - \overline{X}\right) + \beta_2 \left(X_{t-6} - \overline{X}\right) + \beta_3 \left(X_{t-12} - \overline{X}\right) \tag{A.3}$$

$$Y_t = \alpha + \beta_1 \left(X_t - \overline{X}\right) + \beta_2 \left(X_{t-6} - \overline{X}\right) \tag{A.4}$$

$$Y_t = \alpha + \beta_1 \left(X_t - \overline{X}\right) \tag{A.5}$$

where the indoor temperatures Y, is being predicted by the demeaned outdoor temperatures X_{t-k}) lagged by k-hours. The predicted values of these four models are shown in plot A.6

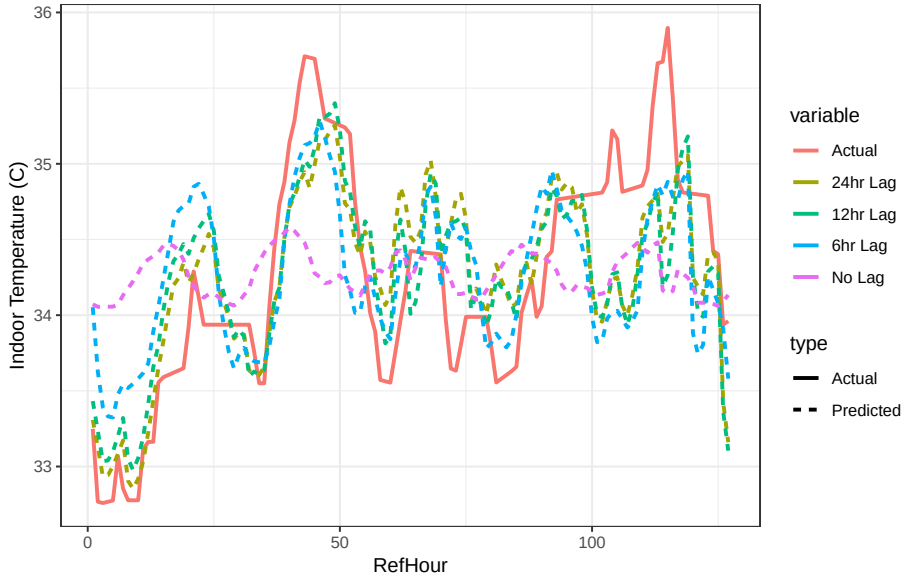


Figure A.6: Predicted Indoor Temperatures

Using these models, we were able to fit a dynamic prediction with the goodness-of-fit metrics located in table A.2. We can see that Equation (A.2), had the best fit with the $R = 0.561$. However, for the sake of integration, it is better to use a more robust model described Equation (A.2), since we can see that the model overfits in Figure A.6 due to the multicollinearity between the terms and this might be an issue for integration. Additional work can be done to validate the model against time-series data from other houses to improve the robustness.

Table A.2: Distributed Lag Model Results

	Dependent variable:			
	InTemp			
	(A.2)	(A.3)	(A.4)	(A.5)
Outdoor Temp (No Lag)	0.058**	0.105***	0.027	0.045**
	(0.025)	(0.020)	(0.017)	(0.021)
Outdoor Temp (24hr Lag)	0.062***			
	(0.020)			
Outdoor Temp (12hr Lag)	0.118***	0.117***		
	(0.020)	(0.020)		
Outdoor Temp (6hr Lag)	0.118***	0.125***	0.147***	
	(0.015)	(0.016)	(0.017)	
Constant	34.096***	34.060***	34.166***	34.242***
	(0.051)	(0.051)	(0.054)	(0.067)
Observations	127	127	127	127
R^2	0.561	0.528	0.401	0.034
Adjusted R^2	0.547	0.516	0.391	0.026
Residual Std. Error	0.508 (df = 122)	0.525 (df = 123)	0.589 (df = 124)	0.744 (df = 125)
F Statistic	39.049*** (df = 4; 122)	45.860*** (df = 3; 123)	41.530*** (df = 2; 124)	4.425** (df = 1; 125)

Note: *p<0.1; **p<0.05; ***p<0.01

A.4 Conclusion and Future Work

To summarize this appendix, we developed the distributed-lag regression model in order to interface with the agent-based model and existing climate model. The intermediate model receives the input from the climate model (i.e. outdoor temperature) and predicts the indoor temperature that the model agents experience. We see that it is possible to reproduce the diurnal cycle using the distributed-lag model. Since most mortality of heat waves occur indoors, the information will be relevant in predicting the health risks. The fully integrated model with the climate and agent-based modules is still being developed.

We outlined an integrated and developed an integrated model for heatwave resilience. Prof. Seth Guikema at the University of Michigan is currently leading multi-institutional efforts to extend the work on finishing and improving the integrated heat wave model with an advanced Agent-Based Model that utilizes social network data.

Appendix B

Network Appendix

B.1 Data and Processing

The analysis was performed on the Aggregate Analysis of ClinicalTrial.gov (AACT) database from the Clinical Trials Transformation Initiative (CTTI) at Duke University [110] and the BioMedTracker Pharma Intelligence Database from Informa LLC. The AACT database was accessed on January 12, 2017 and the Biomed Tracker database was accessed on January 15, 2017.

We used the intersection of the two databases where data existed for the variables of interest. Each distinct trial is defined as a unique national clinical trials (NCT) identifier number. The AACT database provided information on collaboration based on *lead sponsor* and *collaborators*. Hence, we were able to connect each unique NCT number with a set of organizations that were

affiliated with that particular trial to construct our collaboration network. The AACT also provided the official start and end dates of each trial. We selected the trials that were designated as a "drug" interventions as defined by ClinicalTrial.gov. Our analysis is refined to 4,494 organizations[1] and 18,040 trials.

From the BioMedTracker database, we were able to extract the therapeutic research area through by *disease group* designation and FDA approval status of each trial. In our analysis, we considered 21 different disease groups. The disease group data was used to calculate our knowledge mix, research diversification, and collaboration diversity indices which will be further defined in the sections. The trials that were observed are shown in the Figure B.2.

Additional data collection was conducted using a combination of manual web searches and text-mining to classify each actor into 6 actor categories: Academic, Government, Nonprofit, Industry, Hospital System, or Large Pharmaceutical. Table B.1 shows how we categorized each actor. This was done to aid our regression analysis since the roles of different types of organization vary within a collaboration network. For example, academic institutions usually collaborate differently than a nonprofit organization.

[1]In our analysis, we consider each unique name to be a distinct actor. Therefore, subsidiaries that have a different name from the parent company were considered to be separate organizations.

B.2 Constructing the Collaboration Network

Using the *collaborator* section of the AACT database, we structured the data so each clinical trials, which is defined as a unique NCT number is connected to a set of collaborators and lead sponsor. From this structured data, we constructed an undirected 2-mode affiliation network that connected actors (i.e. lead sponsors and collaborators) to a corresponding event (i.e. clinical trial) based on involvement. The 2-mode network is then utilized to construct a 1-mode network using bipartite projection to create an undirected, 1-mode collaboration network. Figure B.1 shows the relationship between a 2-mode and 1-mode network. The one-mode network is represented as a set of actors that are actively conducting clinical trials at that particular time. The set of links exist between pairs of actor that collaborate on the same clinical trials. All our analysis was performed on the 1-mode network.

We built a distinct network for each month which essentially is a "snapshot" of all active clinical trials and their associated collaborations. A clinical trial is considered active in a specific time period if the time period is after the *start date* and before the *end date* as listed in the AACT database. We utilized the Biomed Tracker database to determine whether a trial was suspended (failed) or is associated with a treatment that was by the U.S. Food and Drug Administration (success). The node that represents an actor is removed in the

time-specific network if it is not involved in any active clinical trials. Hence, the number of nodes differ between each time period.

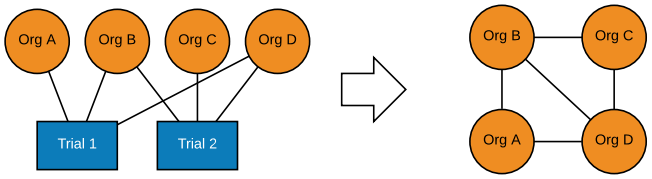


Figure B.1: Bipartite projection of 2-mode affiliation network to a weighted 1-mode collaboration network.

B.3 Definitions of Actor Metrics and Charact

The regression variables are divided into 5 categories: (1) organization type, (2) expertise, (3) structural measures, (4) collaboration measures, and (5) organizational measures. The measured variables that were used in the regression analysis are summarized in Table B.4 in Section B.5.

B.3.1 Organizational Types

In the clinical trials environment, we defined six *organization types*: Academic, Government, Nonprofit, Industry, Hospital System, or Large Pharmaceutical. The organization types of each actor are defined in Table B.1. The organization type designation for each actor is consistent for all time periods. For each of the 4,494, we were able to classify the actors by using publicly available information on the Web. Additionally, we were able to categorize many of the actors based on their registered names in the AACT database. For example, if an organization has "LLC," "Corp.," or "Inc." in their name, we can assume that they belong to industry. Organizations that have "college" or "university" would be classified as academic institutions.

Furthermore, we differentiated between large pharmaceutical companies and industry actors based on their revenue, R & D spending, and profit rankings. The listings for large pharmaceutical companies are listed in Table B.2

Table B.1: Classifications of Organization Types

Category	Description	Count
Academic	university institutions, academic health centers, and learning hospitals	639
Government	national institutes, ministries, and veteran hospitals	123
Nonprofit	patient advocacy, trusts, initiatives, collaborative Groups	398
Industry	pharmaceutical/biotechnology firms (that are not considered large pharmaceutical companies), corporations, multinationals, holdings, and private clinics	2989
Hospital System	healthcare networks and general hospitals	255
Large Pharmaceutical Companies	top 25 companies with highest market capitalization, top 15 revenue, top 15 R&D Budgets from 2016 (See Table B.2 for complete listing)	255

Table B.2: Actors that are considered Large Pharmaceutical Companies

Stryker Orthopaedics	AstraZeneca
Johnson & Johnson	Eli Lilly and Company
Amgen Research Munich GmbH	Johnson & Johnson Pharmaceutical Research & Development, L.L.C.
Alexion Pharma GmbH	Sanofi
Abbott Medical Optics	Wyeth is now a wholly owned subsidiary of Pfizer
Teva Neuroscience, Inc.	Hoffmann-La Roche
Johnson & Johnson Medical, China	Biogen
Astellas Pharma Europe B.V.	Bristol-Myers Squibb
Abbott Products	Amgen
Shire Regenerative Medicine, Inc.	Pfizer
Merck Serono Co., Ltd., Japan	Novartis Pharmaceuticals
Sanofi-Synthelabo	GlaxoSmithKline
Teva Women's Health	Celgene Corporation
Merck Serono S.A., Geneva	Astellas Pharma Inc
Johnson & Johnson Pte Ltd	Astellas Pharma US, Inc.
Teva Pharmaceuticals USA	Bayer
MAP Pharmaceuticals, Inc., a wholly owned subsidiary of Allergan	Genzyme, a Sanofi Company
Stryker Biotech	Abbott
Stryker Instruments	Gilead Sciences
Daiichi Sankyo UK Ltd.	Shire
Sanofi Pasteur MSD	Novartis
Stryker Nordic	Roche Pharma AG
Durata Therapeutics Inc., an affiliate of Allergan plc	Novo Nordisk A/S
Abbott Diabetes Care	Alexion Pharmaceuticals
Roche-Genentech	Merck KGaA
TEVA	Daiichi Sankyo Co., Ltd.
Teva Pharma	Takeda
Abbott Japan Co.,Ltd	Daiichi Sankyo Inc.
Astellas Pharma Global Development, Inc.	Allergan
Allergan Medical	Abbott Vascular
Abbott Diagnostics Division	Teva Pharmaceutical Industries
Astellas Pharma Europe Ltd.	Vertex Pharmaceuticals Incorporated
Celgene	Orthovita d/b/a Stryker
Stryker MAKO Surgical Corp	AbbVie prior sponsor, Abbott
Merck Serono S.A., Switzerland	Stryker Neurovascular
Astellas Pharma China, Inc.	Regeneron Pharmaceuticals
Stryker MAKO Corp	Boehringer Ingelheim
Genentech/Roche	Teva Branded Pharmaceutical Products, R&D Inc.
Johnson & Johnson Medical Companies	Sanofi Pasteur, a Sanofi Company
Bristol Meyers Squibb BMS	AbbVie
Teva Pharmaceutical Industries, Ltd.	Shire Human Genetic Therapies, Inc.
Janssen/GlaxoSmithKline	Janssen, LP
Janssen, GlaxoSmithKline GSK	King Pharmaceuticals is now a wholly owned subsidiary of Pfizer
Astellas Pharma Korea, Inc.	Gamida Cell -Teva Joint Venture Ltd.
Stryker South Pacific	Novartis Vaccines

B.3.2 Expertise

Each trial in the BioMedTracker database is associated with therapeutic area based on the main treatment objectives. In our sample from the BioMedTracker database, there are 21 different therapeutic areas. Figure B.2 shows the distribution of trials with respect to a therapeutic designation.

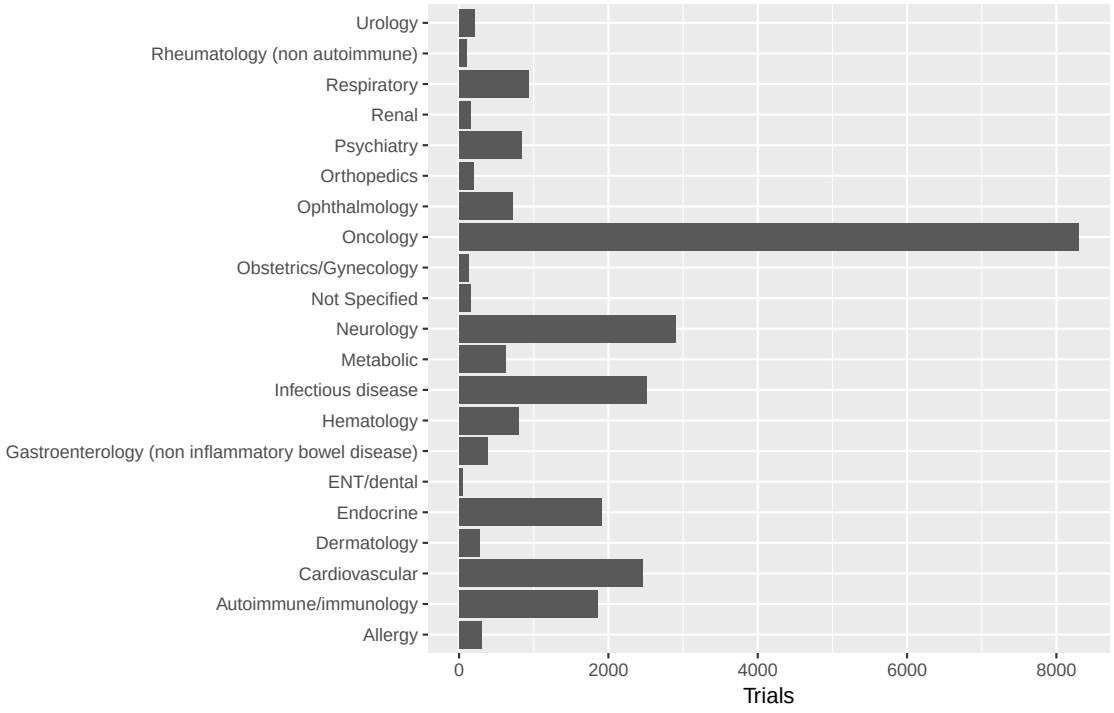


Figure B.2: The distribution of trials in our analysis is shown by therapeutic disease groups from Jan 2006 - Jan 2016

We assumed that the experience and data gained from a clinical trial contributes to the knowledge regarding the therapeutic area. For each actor, we were able to designate them as an expert in a therapeutic area by simply looking at the most trials that the actor was involved with.

B.3.3 Structural Measures

Betweenness Centrality

The *betweenness centrality* BT_{it} for actor i at time t is defined as

$$BT_{it} = \sum_{j,k=1}^{N} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad s.t.\ j \neq k$$

where σ_{ij} represents the shortest path between nodes i and j and $\sigma_{jk}(i)$ represents the total number of shortest paths that goes through node i. This metric is useful for measuring the extent to which a node acts as a bridge between two communities.

The actors that have a high betweenness centrality in the clinical trials collaboration network represents are in positions that would enable them to be a conduit for knowledge flow. Furthermore, these organizations that are in central positions with a high betweenness centrality are also able to tap into the different knowledge bases of a variety of actors. Many large pharmaceutical organizations have a high betweenness centrality. Actors that have low betweenness centrality are typically on the peripheries of a network, such as biotechnology and life science startups.

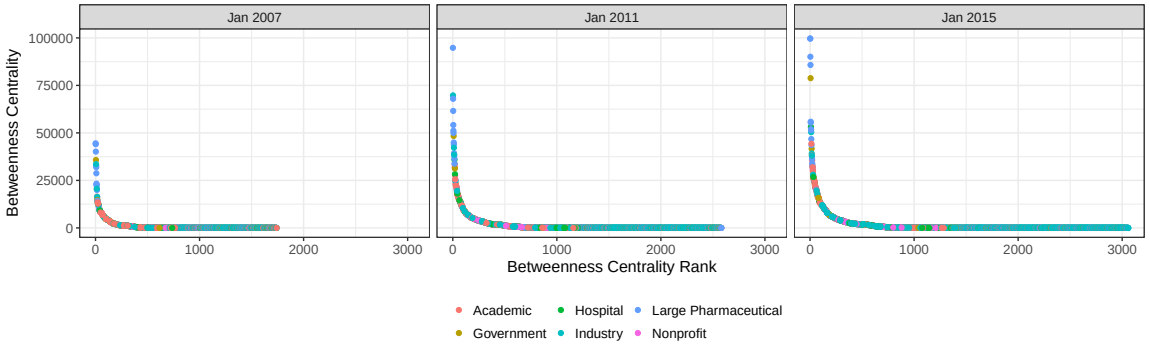


Figure B.3: Rank-size plot for betweenness centrality for Januarys of 2007, 2011, and 2015.

Clustering Coefficient

The *local clustering coefficient* CC_{it} measures the extent to which a node's neighbors all belong to a clique. Formally, the local clustering coefficient for an undirected graph is defined as

$$CC_{it} = \frac{2L_{it}}{\delta_{it}(\delta_{it} - 1)}$$

where L_{it} represents the number of links between the neighbors of actor i at time t, and δ_{it} represents the number of degrees of actor i at time t.

The local clustering coefficient has been shown to be a good indicator of local cohesiveness [36]. The local clustering coefficient also provides us with a measure of how close the local neighborhood of a network is a clique. In the context of collaboration trials, actors that have high measures of local clustering tend speed of knowledge transfer [29]. However, local clustering can

also result in knowledge redundancy in which organizations are essentially stuck in an echo chamber of redundant knowledge.

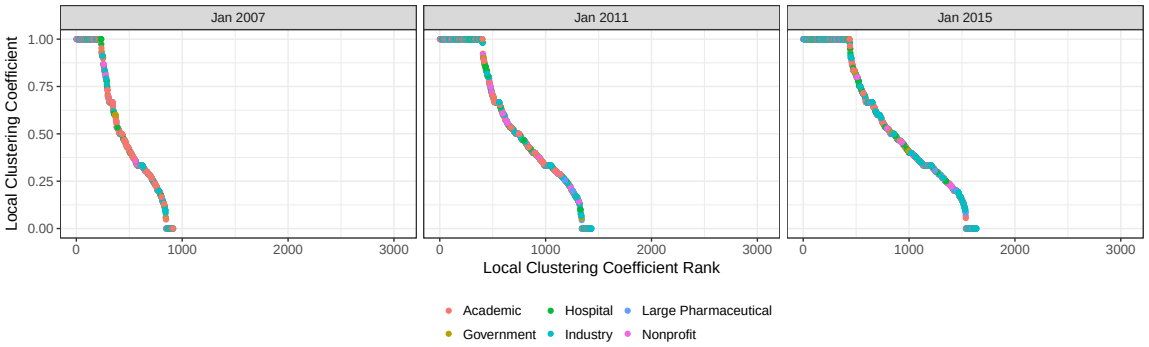


Figure B.4: Rank-size plot for local clustering coefficient for Januarys of 2007, 2011, and 2015. Nodes with less than 2 neighbors are removed.

Degree Centrality

Per conventional definition of degree centrality, this is simply the number of links that are adjacent to the observed node, which we will define as δ_{it}. Also known as degree, δ_{it} is the degree count at time t for actor i.

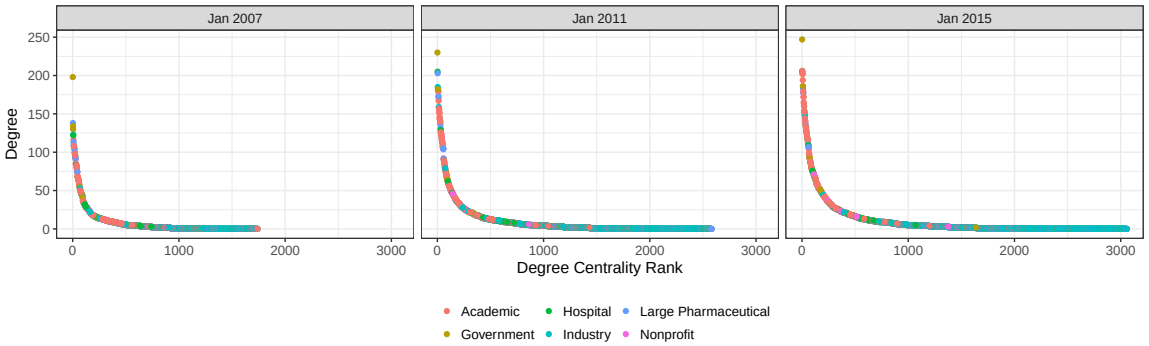


Figure B.5: Rank-size plot for degree centrality for Januarys of 2007, 2011, and 2015. Nodes with less than 2 neighbors are removed.

B.3.4 Organizational Measures

Knowledge Mix

Similar to the knowledge space index was developed by Tomasello et al. [43] to show the distribution of research in different industrial patent areas, our knowledge mix index shows the distribution of an actor's research portfolio in each disease group research area. They considered knowledge to be demonstrated when a company files a patent in a certain industrial area (e.g. aerospace, pharmaceutical, manufacturing), while we considered knowledge mix to be trials that were conducted in each disease group area. We utilize the knowledge mix index to determine the amount of experience gained when an organization conducts research in a particular disease domain. As shown in the main article, the knowledge mix for organization i is defined as a vector $\mathbf{x}_i$ with the element x_{id}. The element x_{id} is defined as

$$x_{id} = \frac{n_{id}}{\sum_{d \in \mathscr{D}} n_{id}} \; \forall i \in \mathscr{F}, \; \forall d \in \mathscr{D} \tag{B.1}$$

The set of firms and therapeutic diseases area are represented respectively as $\mathscr{F}$ and $\mathscr{D}$. If organization i does not have any approvals, then the knowledge mix is a null vector $\mathbf{x}_i = \mathbf{0}$. We define experience (i.e. knowledge) as n_{id} which represents the number of clinical trials in the therapeutic area d that actor i has been involved as a sponsor or partner.

APPENDIX B. NETWORK APPENDIX

Since the knowledge mix is not a scaler, it was not used directly in our regression analysis. However, the knowledge mix is used to calculate research diversification, mean neighbor research diversification, knowledge distance, and collaboration diversity.

Research Diversification

On a firm level, an organization may decide to adopt two approaches: broadly diversify in different therapeutic disease areas (i.e. “jack-of-all-trades”) or specialize in one therapeutic disease area (i.e. “master-of-one”). The decision to diversify is usually driven by the size of the organizations which determines the economies of scope [45]. Larger companies tend to diversify more than smaller companies. Furthermore, the firm must decide whether to collaborate with a jack-of-all-trades or a master-of-one that has a deep understanding in one field.

We quantified *research diversification* using an entropic measure that measures the heterogeneity of actor i's knowledge mix $\mathbf{x}_i$.

$$RD_{it} = \sum_{d \in \mathscr{D}} x_{id} \ln \left(\frac{1}{x_{id}} \right) \tag{B.2}$$

where x_{id} is the element of the knowledge mix vector that represents the percentage of clinical trials experience in disease d. This measure gives us an impression of the level of interdisciplinary in an organization's research portfolio. We assume that a company that has completed 0 trials will have an entropy of 0.

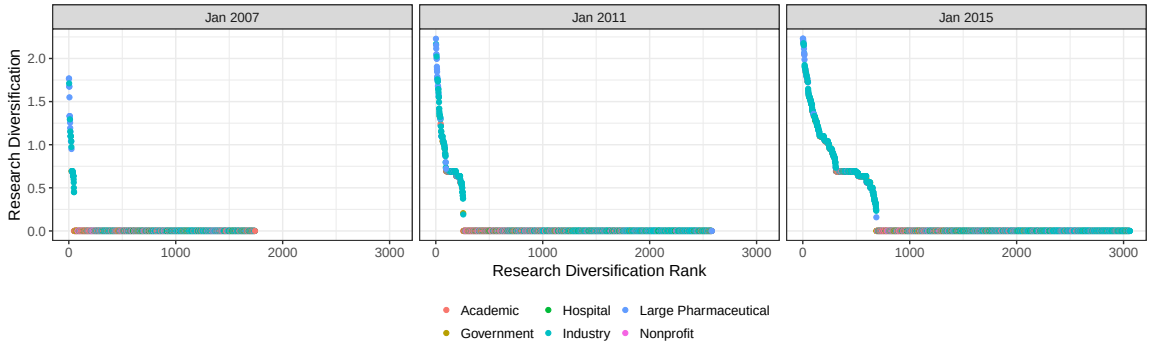


Figure B.6: Rank-size plot for Research Diversification for Januarys of 2007, 2011, and 2015.

B.3.5 Collaboration Measures

Mean Knowledge Distance

Tomasello et al. [43] defined the knowledge distance as the Euclidean distance between organizations i and j at time t. In other economic literature, this is known as the technological distance [44]. This is formally defined as

$$KD_{ijt} = \|\mathbf{x}_j - \mathbf{x}_j\| = \sqrt{\sum_{d \in \mathscr{D}} (x_{id} - x_{jd})}. \tag{B.3}$$

where x_{id} represents an element of the knowledge mix vector $\mathbf{x}_j$ that was defined in B.3.4.

This link-specific metric was meant to compare the differences between two firms' patent portfolios. However, we have adopted this metric to measure the differences between pairs of the organization's research portfolio (i.e. the distribution of trial experience in each therapeutic area). The knowledge distance

is at a maximum ($KD = \sqrt{2}$) when actors are concentrated in two different therapeutic areas. When two firms are concentrated in the same therapeutic area, the knowledge distance equals to 0 because they are identical in expertise.

One of the properties of the Euclidean-based knowledge distance in (B.3) is that the measure takes into account research diversification of an actor. Let's say Actor 1's research portfolio is solely concentrated in Neurology, Actor 2 is solely concentrated in Urology, and Actor 3 is divided between Oncology and Urology. In this situation, the known distance between Actor 1 and Actor 2 is larger than Actor 1 and Actor 3 even despite the fact that both Actor 2 and 3 are in exclusive research fields relative to Actor 1. This is a well-known property of Euclidean Distances and fits our case since we are implying that actors that are more interdisciplinary have more capacity to function in other fields than specialists.

In our analysis, we calculate the mean knowledge distance $\langle KD \rangle_{it}$ for all incident links to actor i at time t, and used it as a variable in our regression. We can define this as

$$\langle KD \rangle_{it} = \frac{\sum_{\{i,j\} \in E(G)} KD_{ij}(t)}{\delta_i(t)} \quad s.t.\ i \neq j$$

where δ_i is the number of degrees for actor i.

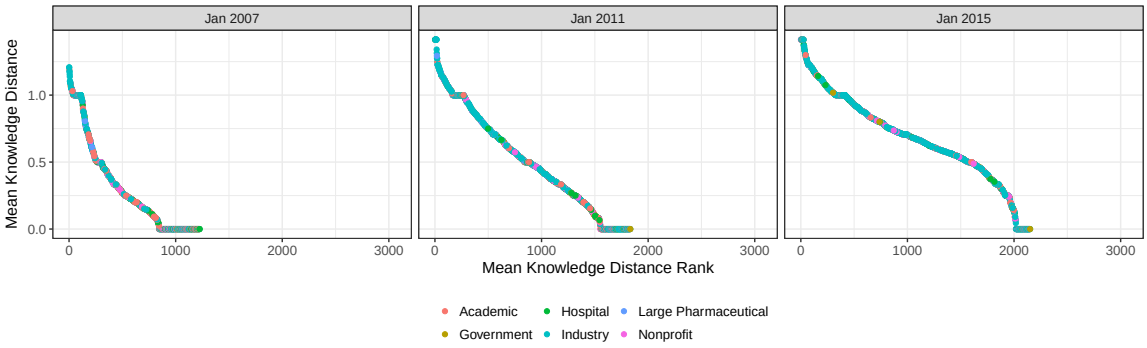


Figure B.7: Rank-size plot for Mean Knowledge Distance for Januarys of 2007, 2011, and 2015.

Collaboration Diversity

For each actor, the *collaboration diversity* measures how interdisciplinary the collaborators are in a given instance. Specifically, we determine how many experts in each field an actor is collaborating with and how heterogeneous that distribution is.

By simply observing which element of the knowledge mix vector in Section B.3.4 that has the largest value, we can determine the therapeutic area that an actor would be considered an "expert."[2] Based on this method, we were able to designate each node in our graph with an expertise and calculate the level of diversity of collaborators for each actor. This measure is similar to the entropic measure in Equation (B.2).

For each actor i, there exist a set of collaborators (neighbors) $\mathscr{N}_i$ with a corresponding vector of expertise count $\mathbf{z}_i(\mathscr{N}_i)$. The element $z_{id} \in \mathbf{z}_i$ is equal to number of collaborators that are experts in therapeutic field d. We can formally define our collaboration diversity as

$$CD_{it} = \sum_{d \in \mathscr{D}} z_{id} \ln\left(\frac{1}{z_{id}}\right) \tag{B.4}$$

For actors that have no collaborators, we assume $CD = 0$.

[2]For some of the actors, they have expertise in 2 or more therapeutic areas (e.g. $\max_{d \in \mathscr{D}} x_{id} = x_{i1} = x_{i2}$). In these cases, we designate the actors as a separate group from the rest of the therapeutic designations. Partners with no expertise in any therapeutic areas are excluded.

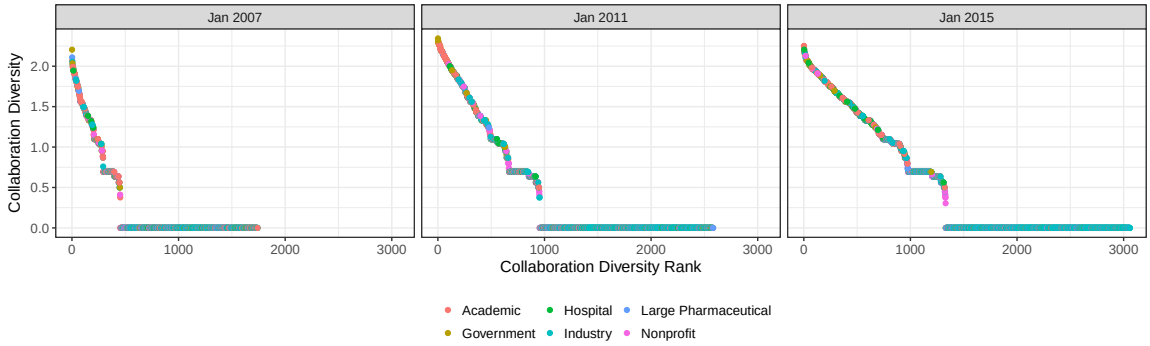


Figure B.8: Rank-size plot for Collaboration Diversity for Januarys of 2007, 2011, and 2015.

Mean Neighbor Research Diversification

From the research diversification that is defined in Equation (B.2), we can determine mean research diversification of all the neighboring organization of actor i at any given time period t. The mean research diversification $\langle RD \rangle_{it}$ is simply

$$\langle RD \rangle_{it} = \frac{\sum_{\{i,j\} \in E(G)} RD_{ij}(t)}{\delta_i(t)}$$

where δ_i is the number of degrees for node i.

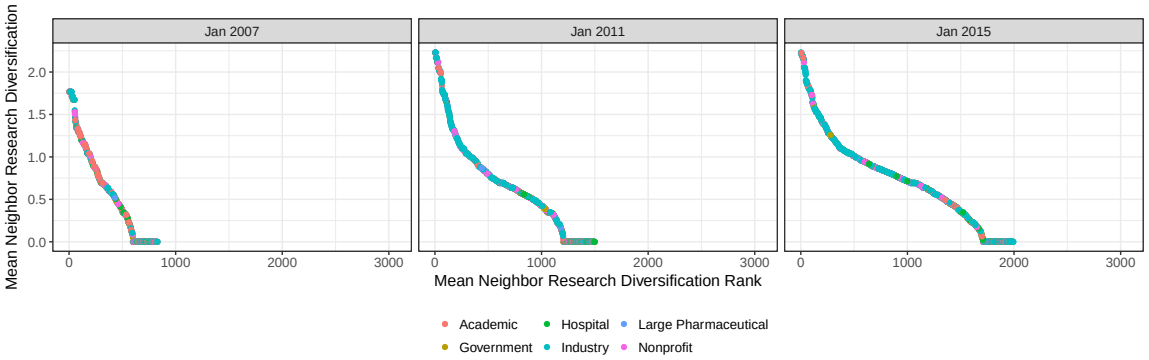


Figure B.9: Rank-size plot for Mean Neighbor Research Diversification for Januarys of 2007, 2011, and 2015.

B.4 Fitting Regression Models

Using the set of 121 temporal collaboration networks that was constructed, we created a panel dataset composed of the network measures in Section B.3.3. We plotted the trends for all of the months in Figure 2 in the main article. For our regression analysis, we discretized the data based on 6-month intervals, starting in January 2006 and ending in January 2016 (19 time periods).

In order to obtain explanatory relationships of each network measure on treatment development success, we respectively utilized a negative-binomial generalized linear model (GLM) and beta regression model to predict two sets of lagged dependent variables that represent research output and productivity: cumulative trial successes and trial success rate. Given that cumulative trial successes are a count variable, we selected the negative binomial regression was selected to be our model. We selected negative binomial regression for

the cumulative trial success over Poisson regression due to the over-dispersion of the response variable. For the trials success rate, the beta regression was selected to predict a response variable with a continuous unit interval (0,1).

We also considered 3 different lag lengths: 1, 2, and 5 years to analyze time-scale effects. For each dependent variable, we developed the main model by heuristically eliminating variables based on the variational inflation factor (VIF) and P-value of each regression estimation. The main model was fitted for all 3 lag durations. The GLM and beta regression models are defined in Equations (B.5) and (B.6) respectively, and referenced in the main article. We have found that the main model is the most useful since it only includes variables that have significant explanatory power and reduced multicollinearity. In addition to the main model, we also developed a base model with only control variables and a comprehensive model with all variables. The results for the base model (control variables), main model (selected variables), and comprehensive model (all variables) are shown in Tables B.5 - B.7 and B.9 - B.11. We also included

To account for confounding temporal factors (e.g. regulations, policy changes), we added fixed effects for each time period. We also included fixed effects for each organizational class (e.g. academic, nonprofit) to capture the differences in actors' behaviors.

The GLM and beta Regression will have the following form

$$g(y_{ijt}) = \beta_0 + \beta_1 x_{1ij(t-k)} + \beta_2 x_{2ij(t-k)} + \ldots + \beta_{Mij(t-k)} + \gamma_t + \kappa_j + \epsilon_{ijt}$$

where $g(\cdot)$ is the link function and γ_t is the fixed effect of each 6-month time interval between 2006-2016. There are M covariates (independent variables) designated as x_{ijk} that measures an attribute of actor i of organization type j at time t. We employed the statistical software R to estimate the coefficients β using the method of maximum likelihood to regress against the response variable y_{ijt}.

B.4.1 Cumulative Trial Successes Regression

We utilized a negative binomial regression class of GLMs to estimate the linear relationship of cumulative trial successes CTS_{ijt} for all actors i of organization type j and time t. Our main model is defined as

$$\begin{aligned}\log(CTS_{it}) = \beta_0 + \beta_1 PrevSucc_{i(t-k)} + \beta_2 Trials_{i(t-k)} + \beta_3 CD_{i(t-k)} + \beta_4 \langle KD \rangle_{i(t-k)} \\ + \beta_5 CC_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it}\end{aligned} \tag{B.5}$$

The variables that were selected for this model include attributes that depict network, collaboration and organization characteristics that gave us the highest

cross-validation predictive accuracy using the method of random holdouts. These covariates include previous success dummy $PrevSucc_{i(t-k)}$, cumulative trials conducted $Trials_{i(t-k)}$, collaboration diversity $CD_{i(t-k)}$, mean knowledge distance $\langle KD \rangle_{i(t-k)}$, and local clustering coefficient $CC_{i(t-k)}$. We standardized cumulative trials conducted, collaboration diversity, mean knowledge distance, and local clustering coefficients in the main model.

Since the dependent variable is lagged, all the independent variables are taken from an earlier period in our panel dataset with lag differences $k =$ 1, 2, and 5 years. The dummy variables γ_t and κ_i are the fixed effects of time and organizational class per Table B.1. All non-dummy variables were standardized to control for varying magnitudes. For more information on these variables, refer to Section B.3. The coefficients for (B.5) are listed as Models A1-2, A2-2, and A3-2 in Tables B.5 - B.7.

It is possible that there exists a reverse causality in which the network structure is the result of the organizational success of the company which makes the network metrics endogenous. For instance, organizations may want to work with a more successful company (i.e., preferential attachment), resulting in more trials. We attempted to control the endogenous effects by including past success as a dummy variable $PrevSucc_{i(t-k)}$ in order to account for confounding factors between network structure and cumulative success in trials. In our model, $PrevSucc_{t-k} = 1$ when a company has participated in at least one successful

clinical trial at or before time $t-k$, otherwise $PrevSucc = 0$.

To illustrate the robustness of our model to explain cumulative trial successes, two other set of regression models were developed in addition to (B.5) to explain the relative effects of all the variables. The base models A1-1, A2-1, and A3-1 only include the control variables while the comprehensive models A1-3, A2-3, and A3-3 include all the variables. Along with the standardized variables in the main model, betweenness centrality and collaboration diversity were also standardized in the comprehensive model. The 3 sets of models for each time lag (1 year, 2 years, and 5 years) resulted in a total of 9 models with results in Tables B.5 - B.7. We also included a comparison of these models using a likelihood ratio test in Table B.8 since the three models are nested within each other. The results show that most of our selected variable's p-values in the main model for all three lag sizes are significant to at least $P < .01$ with most of them being significant to $P < .001$. All VIFs are less than 3 for all models. Previous success has the largest positive effect and that effect size was inversely related to lag sizes. Mean knowledge distance has a negative effect for 1 and 2 years lags, while local clustering coefficient had a negative effect for all lag sizes.

There is evidence of overdispersion in our data since overdispersion parameter θ is greater than 2.2 for 1 and 2-year lags. There is less overdispersion for the 5 year lag with $\theta > 1.1$. We also assume linearity in our model since we

do not observe any distinct higher-order polynomial relationship between the response and explanatory variables.

B.4.2 Trial Success Rate Regression

The trial success rate was analyzed using beta regression for lags of 1, 2, and 5 years. Trials success rate SR_{it} is defined as the cumulative number of successful clinical trials divided by the total number of trials conducted up to time t for actor i of organization type j.

$$SR_{it} = \frac{CTS_{it}}{Trials_{it}}$$

Since the Trials Success Rate in our panel dataset includes 1 and 0 values which are excluded in a beta regression, we transformed the SR into the dependent variable to exclude those extreme values with the function, $\frac{y \cdot (n-1) + .5}{n}$ as suggested by [111].

The resulting model is

$$\begin{aligned}\text{logit}(SR_{it}) = \beta_0 + \beta_1 PrevExp_{i(t-k)} + \beta_2 \langle KD \rangle_{i(t-k)} + \beta_3 CC_{i(t-k)} + \beta_4 RD_{i(t-k)} \\ + \beta_5 \langle RD \rangle_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it}\end{aligned} \tag{B.6}$$

For the beta regression model, mean knowledge distance $\langle KD \rangle_{i(t-k)}$, clustering

coefficient $CC_{i(t-k)}$, research diversification $RD_{i(t-k)}$, and mean neighbor's research diversification $\langle RD \rangle_{i(t-k)}$ were standardized. The dummy variable $PrevExp_{i(t-k)}$ is meant to capture whether the organization has previous experience in clinical trials. Specifically, $PrevExp_{i(t-k)}$ is equal to 1 if the actor has conducted more than 6 trials[3] before time $t-k$. For the beta regression models, we also controlled for time using fixed effect dummy variables γ_t. The fixed effects for each organizational class (defined in Table B.1) are captured in κ_i. For our OLS models, we assumed the link function is an identity function per convention. The main models with the selected variables defined in Equation (B.6) are listed as Model B1-2, B2-2, and B3-2 in Tables B.9 - B.11.

As we did for the analysis of cumulative trial successes, we developed 2 sets of beta regression models using control variables and all variables to show robustness. The base beta regression models with only the control variables are listed as base models B1-1, B2-1, and B3-1. The comprehensive models B1-3, B2-3, and B3-3 show the beta regression with all the interesting variables. The results are shown in Tables B.9 - B.11. In our base model, most variables with significance of least $P < .001$. The previous experience dummy variable was not significant for a 5-year lag. The local clustering coefficient was the only variable with a negative effect for all lag sizes.

The dependent variable in our data follows a beta distribution. Our data

[3]we used 6 because it is the average number of trials

is relatively noisy therefore we are assuming that the response variable and explanatory variables have a linear relationship which is consistent with the assumptions of a beta regression.

B.5 Supplementary Tables and Figures

Table B.3: Summary Statistics of Variables (Unstandardized)

Statistic	N	Mean	St. Dev.	Min	Max
Cumulative Trial Success	9,775	2.684	9.610	0	187
Trial Success Rate	9,775	0.283	0.369	0.000	1.000
Cumulative Trials Conducted	9,775	5.451	19.044	0	398
Mean Knowledge Distance	9,775	0.768	0.272	0.000	1.414
Local Clustering Coefficient	9,775	0.441	0.294	0.000	1.000
Vertex Degree	9,775	27.673	39.929	2	258
Research Diversification	9,775	0.334	0.536	0.000	2.262
Mean Neighbor Research Diversification	9,775	0.706	0.438	0.000	2.262
Betweenness Centrality	9,775	4,815.421	9,910.420	0.000	110,661.500
Collaborator Diversity	9,775	1.100	0.676	0.000	2.344

Table B.4: Summary of Variables

Variable	Description	Units	Source
Dependent Variables			
Cumulative Trial Success	Number of trials (phases) conducted up to time t that led to an FDA approved treatment	trials	BiomedTracker
Trial Success Rate	Rate of trials (phases) that led to an FDA approved treatment at time t	%	BiomedTracker
Independent Variables (Lagged)			
Mean Knowledge Distance	The mean of all incident edge's (i.e. collaborations) knowledge distances to the actor	n/a	AACT
Local Clustering Coefficient	Measures the degree to which a node's neighbors are a clique	n/a	AACT
Vertex Degree	The number edge's (collaborations) an actor is involved with	links	AACT
Research Diversification	Entropic measure of research mix (portfolio)	n/a	AACT/BiomedTracker
Mean Neighbor Research Diversification	Mean value of all neighbor collaborator's research diversity	n/a	AACT/BiomedTracker
Betweenness Centrality	Network measure determining the extent that a node is a bridge	n/a	AACT
Cumulative Trials Conducted	The number of trials that an actor has conducted up up to time t	trials	AACT
Collaborator Diversity	The diversity of collaborators with respect to their specialization	n/a	AACT/BiomedTracker
Previous Success	Actor has participated in at least one trial that led to an FDA approved treatment	0/1	BiomedTracker
Previous Experience	Actor has experience in conducting at least six trials	0/1	AACT

Table B.5: Negative Binomial Regression on Cumulative Trials Success (1-year lag)

	Cumulative Trials Success		
	(Model A1-1)	(Model A1-2)	(Model A1-3)
	Control Variables	Selected Variables	All Variables
Previous Success	2.665***	2.709***	2.572***
	(0.036)	(0.036)	(0.038)
Cumulative Trials Conducted	0.479***	0.331***	0.290***
	(0.008)	(0.007)	(0.009)
Collaboration Diversity		0.240***	0.182***
		(0.015)	(0.017)
Mean Knowledge Distance		−0.330***	−0.248***
		(0.016)	(0.018)
Local Clustering Coef.		−0.075***	−0.064***
		(0.014)	(0.014)
Mean Neighbor Research Diversification			0.041***
			(0.013)
Betweenness Centrality			0.015
			(0.011)
Research Diversification			0.135***
			(0.015)
Constant	−2.317***	−2.623***	−2.460***
	(0.135)	(0.130)	(0.130)
Observations	9,775	9,775	9,775
Log Likelihood	−11,613.780	−11,256.750	−11,212.030
θ	2.651*** (0.090)	3.419*** (0.131)	3.635*** (0.143)
Akaike Inf. Crit.	23,279.560	22,571.500	22,488.050

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table B.6: Negative Binomial Regression on Cumulative Trials Success (2-year lag)

	Cumulative Trials Success		
	(Model A2-1)	(Model A2-2)	(Model A2-3)
	Control Variables	Selected Variables	All Variables
Previous Success	2.113***	2.177***	2.029***
	(0.030)	(0.031)	(0.035)
Cumulative Trials Conducted	0.438***	0.301***	0.251***
	(0.009)	(0.008)	(0.010)
Collaboration Diversity		0.270***	0.213***
		(0.015)	(0.017)
Mean Knowledge Distance		−0.263***	−0.194***
		(0.017)	(0.018)
Local Clustering Coef.		−0.084***	−0.074***
		(0.014)	(0.014)
Mean Neighbor Research Diversification			0.056***
			(0.014)
Betweenness Centrality			0.026**
			(0.013)
Research Diversification			0.128***
			(0.015)
Constant	−1.531***	−1.877***	−1.723***
	(0.116)	(0.111)	(0.112)
Observations	9,131	9,131	9,131
Log Likelihood	−11,870.990	−11,578.300	−11,536.560
θ	2.230*** (0.081)	2.850*** (0.116)	2.987*** (0.125)
Akaike Inf. Crit.	23,789.990	23,210.600	23,133.110

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table B.7: Negative Binomial Regression on Cumulative Trials Success (5-year lag)

	Cumulative Trials Success		
	(Model A3-1) Control Variables	(Model A3-2) Selected Variables	(Model A3-3) All Variables
Previous Success	1.390***	1.285***	1.127***
	(0.044)	(0.046)	(0.051)
Cumulative Trials Conducted	0.321***	0.236***	0.092***
	(0.015)	(0.015)	(0.021)
Collaboration Diversity		0.269***	0.186***
		(0.021)	(0.022)
Mean Knowledge Distance		0.045**	0.092***
		(0.020)	(0.021)
Local Clustering Coef.		−0.154***	−0.128***
		(0.019)	(0.019)
Mean Neighbor Research Diversification			0.068***
			(0.018)
Betweenness Centrality			0.125***
			(0.022)
Research Diversification			0.159***
			(0.022)
Constant	−0.692***	−0.797***	−0.743***
	(0.093)	(0.092)	(0.092)
Observations	6,094	6,094	6,094
Log Likelihood	−9,527.913	−9,378.134	−9,335.782
θ	1.115*** (0.043)	1.289*** (0.053)	1.338*** (0.056)
Akaike Inf. Crit.	19,091.830	18,798.270	18,719.560

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table B.8: Likelihood Ratio Test for Cumulative Trials Success Regression Models

Lag	Model	theta	Resid. df	2 x log-lik.	Test	df	LR stat.	Pr.Chi.
1 year	A1-1	2.65	9749	-23225.56				
1 year	A1-2	3.42	9746	-22511.50	A1-1 vs A1-2	3	714.06	0.00
1 year	A1-3	3.63	9743	-22422.05	A1-2 vs A1-3	3	89.45	0.00
2 year	A2-1	2.23	9107	-23739.99				
2 year	A2-2	2.85	9104	-23154.60	A2-1 vs A2-2	3	585.39	0.00
2 year	A2-3	2.99	9101	-23071.11	A2-2 vs A2-3	3	83.49	0.00
5 year	A3-1	1.12	6076	-19053.83				
5 year	A3-2	1.29	6073	-18754.27	A3-1 vs A3-2	3	299.56	0.00
5 year	A3-3	1.34	6070	-18669.56	A3-2 vs A3-3	3	84.70	0.00

Table B.9: Beta Regression on Trials Success Rate (1-year lag)

	Trials Success Rate		
	(Model B1-1) Control Variables	(Model B1-2) Selected Variables	(Model B1-3) All Variables
Previous Experience	0.404^{***}	0.236^{***}	0.238^{***}
	(0.043)	(0.052)	(0.053)
Mean Knowledge Distance		0.025^{*}	0.028^{**}
		(0.014)	(0.014)
Local Clustering Coef.		-0.037^{***}	-0.041^{***}
		(0.014)	(0.015)
Research Diversification		0.092^{***}	0.097^{***}
		(0.019)	(0.020)
Mean Neighbor Research Diversification		0.100^{***}	0.102^{***}
		(0.014)	(0.014)
Betweenness Centrality			0.0003
			(0.018)
Collaboration Diversity			-0.020
			(0.017)
Constant	-0.533^{***}	-0.434^{***}	-0.422^{***}
	(0.146)	(0.147)	(0.147)
Observations	9,775	9,775	9,775
R^2	0.037	0.052	0.053
Log Likelihood	36,126.930	36,171.430	36,172.170

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table B.10: Beta Regression on Trials Success Rate (2-year lag)

	Trials Success Rate		
	(Model B2-1) Control Variables	(Model B2-2) Selected Variables	(Model B2-3) All Variables
Previous Experience	0.412***	0.196***	0.196***
	(0.049)	(0.061)	(0.061)
Mean Knowledge Distance		0.034**	0.038***
		(0.014)	(0.014)
Local Clustering Coef.		−0.041***	−0.044***
		(0.015)	(0.015)
Research Diversification		0.103***	0.109***
		(0.020)	(0.021)
Mean Neighbor Research Diversification		0.115***	0.118***
		(0.014)	(0.014)
Betweenness Centrality			0.006
			(0.019)
Collaboration Diversity			−0.029
			(0.018)
Constant	−0.787***	−0.671***	−0.658***
	(0.116)	(0.117)	(0.118)
Observations	9,131	9,131	9,131
R^2	0.037	0.057	0.057
Log Likelihood	32,728.370	32,782.770	32,784.110

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table B.11: Beta Regression on Trials Success Rate (5-year lag)

	Trials Success Rate		
	(Model B3-1) Control Variables	(Model B3-2) Selected Variables	(Model B3-3) All Variables
Previous Experience	0.362^{***}	0.077	0.055
	(0.083)	(0.105)	(0.107)
Mean Knowledge Distance		0.031^{*}	0.038^{**}
		(0.018)	(0.018)
Local Clustering Coef.		-0.040^{**}	-0.041^{**}
		(0.018)	(0.018)
Research Diversification		0.102^{***}	0.100^{***}
		(0.026)	(0.027)
Mean Neighbor Research Diversification		0.110^{***}	0.113^{***}
		(0.017)	(0.018)
Betweenness Centrality			0.033
			(0.024)
Collaboration Diversity			-0.040^{*}
			(0.021)
Constant	-0.822^{***}	-0.750^{***}	-0.740^{***}
	(0.087)	(0.088)	(0.089)
Observations	6,094	6,094	6,094
R^2	0.033	0.050	0.051
Log Likelihood	18,953.220	18,986.650	18,988.750

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Figure B.10: Correlation matrix for the network, organizational, and collaboration measures over the 2006-2016 time frame at 6-month intervals with one year lag for Cumulative Trial Success and Trial Success Rate. (N = 9055).

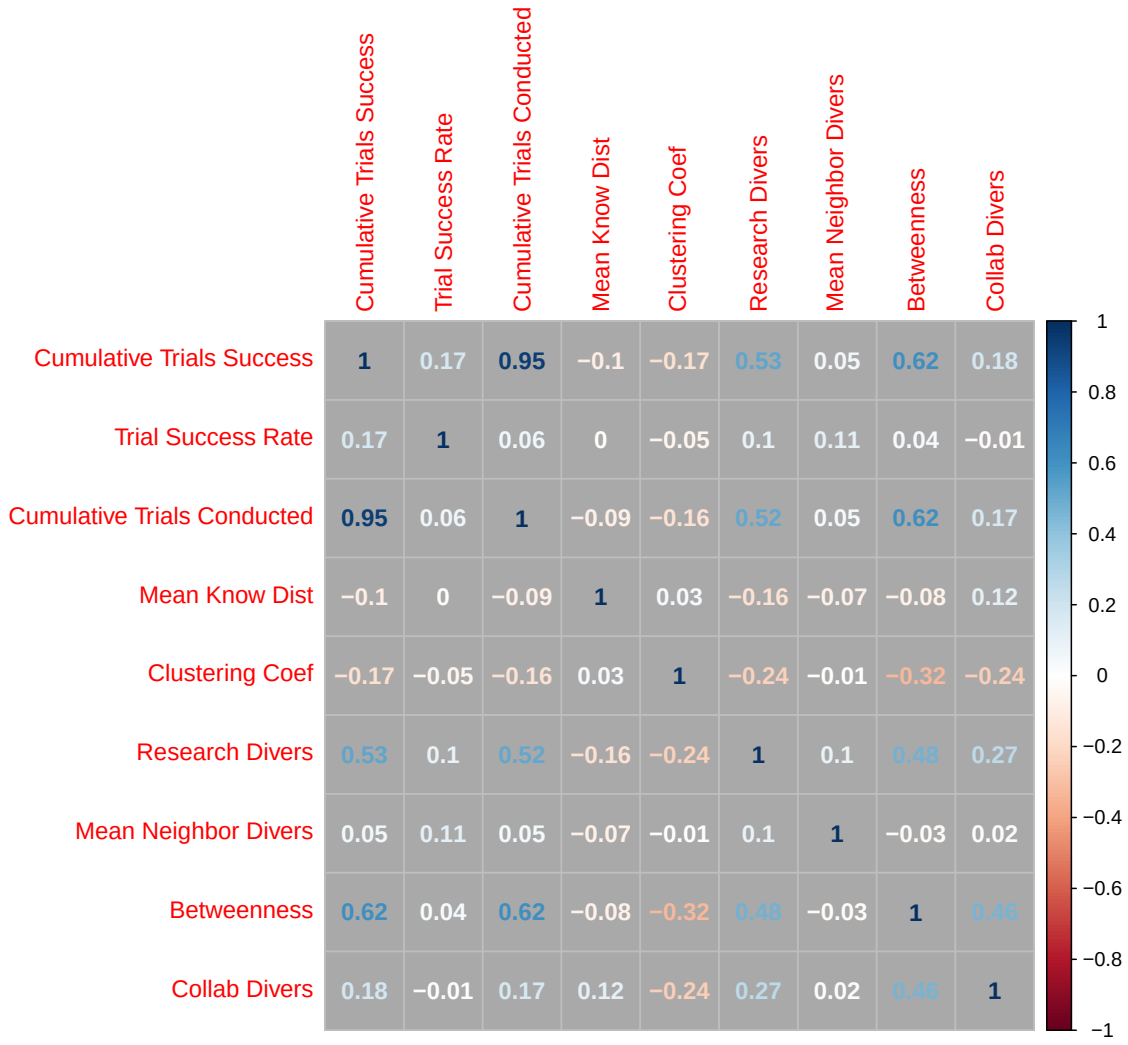


Figure B.11: The scatterplot comparing collaboration diversity and local clustering coefficient for each actor. Large pharamceutical companies and other actors are distinguished by shapes. The color gradient is scaled based on cumulative trial successes.

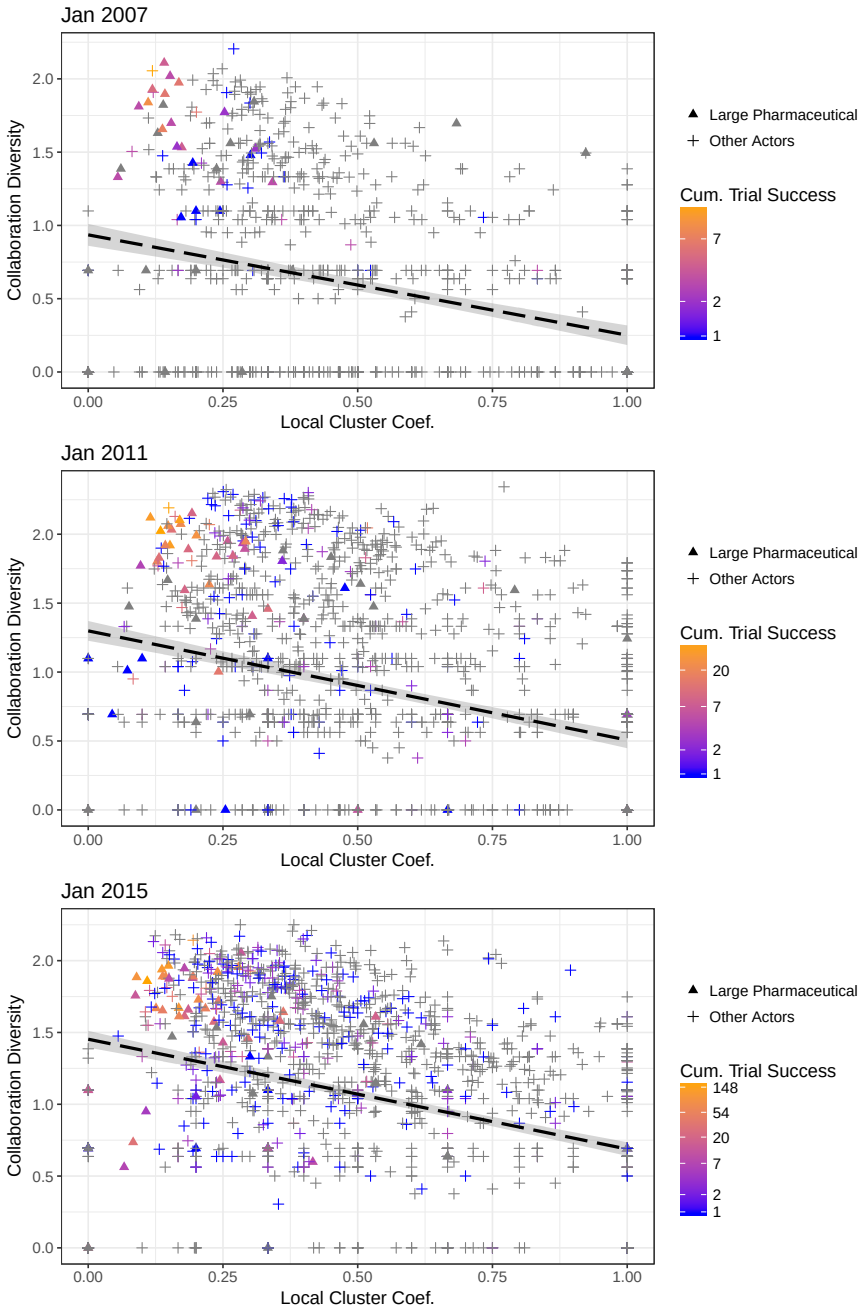
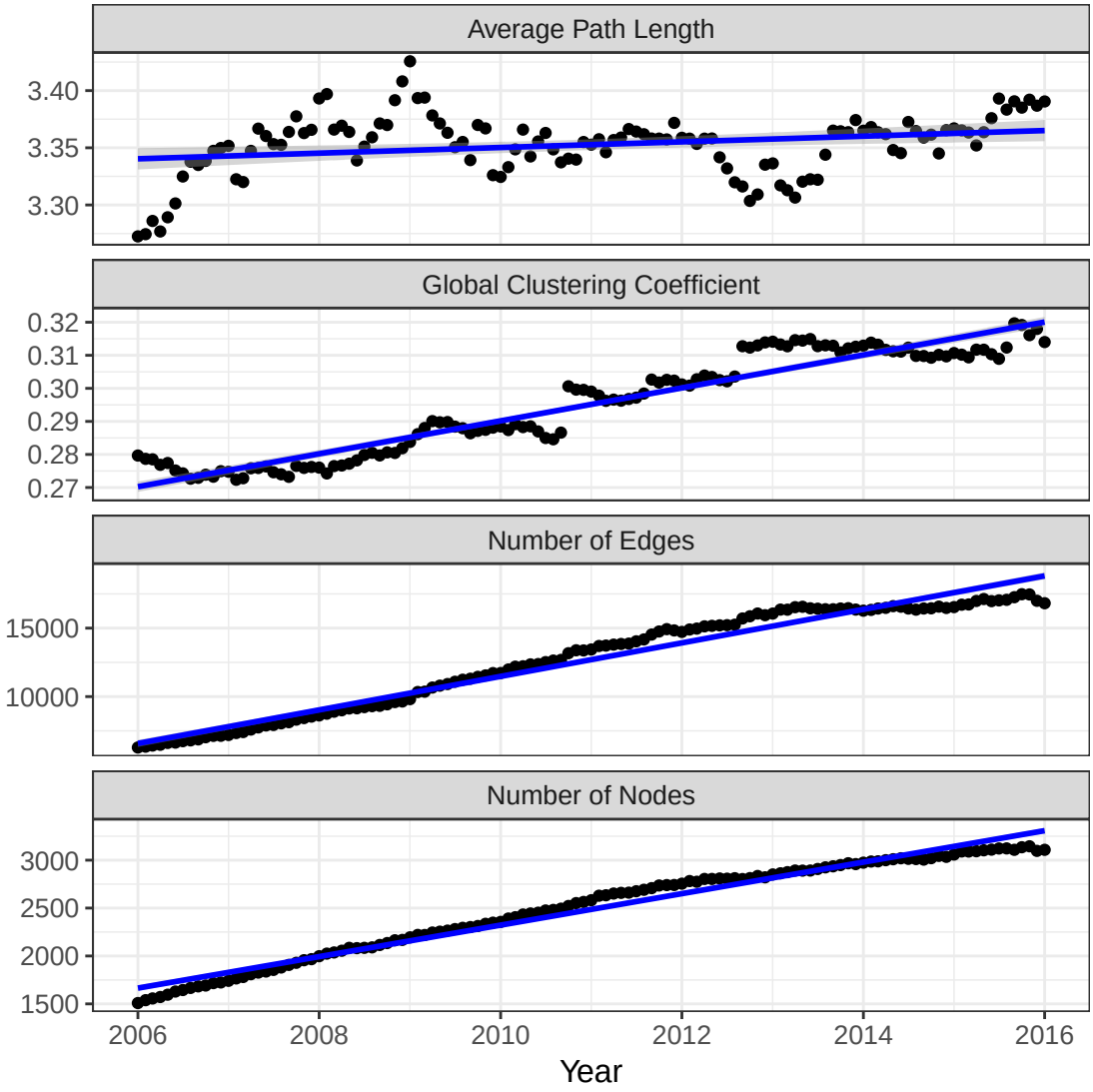


Figure B.12: Global characteristics of the network from 2006 to 2016. The black dots represent the calculated network metric while the blue line represents a linear trendline.



Appendix C

Equations for a multi-component global population model

C.1 Population Equations and Variables

Sets

$$\text{Sex}: \mathscr{S} = \{\text{Male}, \text{Female}\}$$

$$\text{Age Group}: \mathscr{A} = \{\text{Age Group 1}, \ldots, \text{Age Group 4}\}$$

$$\text{Socioeconomic Group}: \mathscr{K} = \{\text{Rich}, \text{Poor}\}$$

Equations

$$\text{Age Group 1 (0-14 years)}: \quad \frac{\mathrm{d}P_{1jr}(t)}{\mathrm{d}t} = B_{jr}(t) + P_{1jr}(t) \cdot \big[- MR_{12}(t) \big] - \sum_{k \in \mathscr{K}} DR_{1jkr}(t) \cdot P_{1jkr}(t) \tag{C.1}$$

$$\text{Age Group 2 (15-49 years)}: \quad \frac{\mathrm{d}P_{2jr}(t)}{\mathrm{d}t} = P_{2jr}(t) \cdot \big[MR_{12}(t) - MR_{23}(t) \big] - \sum_{k \in \mathscr{K}} DR_{2jkr}(t) \cdot P_{2jkr}(t) \tag{C.2}$$

$$\text{Age Group 3 (50-64 years)}: \quad \frac{\mathrm{d}P_{3jr}(t)}{\mathrm{d}t} = P_{3jr}(t) \cdot \big[MR_{23}(t) - MR_{34}(t) \big] - \sum_{k \in \mathscr{K}} DR_{3jkr}(t) \cdot P_{3jkr}(t) \tag{C.3}$$

$$\text{Age Group 4 (65+ years)}: \quad \frac{\mathrm{d}P_{4jr}(t)}{\mathrm{d}t} = P_{4jr}(t) \cdot MR_{34}(t) - \sum_{k \in \mathscr{K}} DR_{4jkr}(t) \cdot P_{4jkr}(t) \tag{C.4}$$

$$\text{Total Fertility Rate}: \quad TFR_r(t) = C_{\mathrm{M}r}(t) \cdot C_{\mathrm{C}r}(t) \cdot TNM \tag{C.5}$$

$$\text{Marriage Index}: \quad C_{\mathrm{M}r}(t) = \alpha_{\mathrm{M}} + \beta_{\mathrm{EM}} \cdot \big(1 - EF_r(t)\big) \tag{C.6}$$

$$\text{Contraception Index}: \quad C_{\mathrm{C}kr}(t) = \alpha_{\mathrm{C}} + \beta_{\mathrm{EC}} \cdot \big(1 - EF_r(t)\big) + \beta_{\mathrm{HC}} \cdot \big(1 - HF_{kr}(t)\big) \tag{C.7}$$

APPENDIX C. EQUATIONS FOR A MULTI-COMPONENT GLOBAL POPULATION MODEL

$$\textbf{General Fertility Rate}: GFR_r(t) = \beta_1 \cdot TFR_r(t) \tag{C.8}$$

$$\textbf{Male Births}: B_{\mathrm{M}r}(t) = s_{\mathrm{M}} \cdot P_{2Fr}(t) \left(\frac{GFR_r(t)}{1000}\right) \tag{C.9}$$

$$\textbf{Female Births}: B_{\mathrm{F}r}(t) = (1 - s_{\mathrm{M}}) \cdot P_{2Fr}(t) \left(\frac{GFR_r(t)}{1000}\right) \tag{C.10}$$

$$\textbf{Total Population}: TP_r(t) = \sum_{i \in \mathscr{A}} \sum_{j \in \mathscr{S}} \sum_{k \in \mathscr{K}} P_{ijr}(t) \tag{C.11}$$

$$\textbf{Death Rate}: DR_{ijkr}(t) = MDR_{ijkr} + DF_{ijkr}(t) + DH_{ijkr}(t) \tag{C.12}$$

$$\textbf{Food Shortage Death Rate}: DF_{ijkr}(t) = \omega_{\mathrm{F}ijkr} \cdot \max\left\{\left(\frac{\widehat{NR}(t) - \widehat{NC}_{ijkr}(t)}{\widehat{NR}(t)}\right),\ 0\right\} \tag{C.13}$$

$$\textbf{Healthcare Shortage Deaths Rate}: DH_{ijkr}(t) = \omega_{\mathrm{H}ijkr} \cdot \big(1 - HA_{kr}(t)\big) \tag{C.14}$$

Table C.1: Variables for Population Submodel

Variables	Description	Type
B_{jr}	Birth count in for sex j and socioeconomic group k in region r	Endogenous
$C_{\mathrm{C}k}$	Bongaart's fertility index for contraception for socioeconomic group k in region r	Endogenous
$C_{\mathrm{M}r}$	Bongaart's fertility index for marriage for socioeconomic group k in region r	Endogenous
DF_{ijkr}	Food shortage impact on deaths for age group i, sex j, socioeconomic group k in region r	Endogenous
DH_{ijkr}	Health service shortage impact on deaths for age group i, sex j, and socioeconomic group k in region r	Endogenous
DR_{ijkr}	Death rate in age group i, sex j, and socioeconomic group k in region r	Endogenous
EF_r	Female access to education in region r	Endogenous
$\widehat{NC}_{ijkr}$	Nutrition consumption per capita for age group i, sex j, and socioeconomic group k in region r	Endogenous
$\widehat{NR}$	Nutrition requirement per capita	Endogenous
GFR_r	General fertility rate in region r	Endogenous
HA_{kr}	Socioeconomic group k's access to health care in region r	Endogenous
HF_r	Females' access to health care in region r	Endogenous
MDR_{ijkr}	Minimum death rate for age group i, sex j, socioeconomic group k in region r	Constant
$MR_{(i-1)ij}$	Maturation rate from age $i-1$ to i	Constant
P_{ijkr}	Population count in age group i, sex j, and socioeconomic group k in region r	Endogenous
s_{M}	Male sex ratio	Constant
TF	Total fecundity	Constant
TFR_r	Total fertility rate in region r	Endogenous
TNM	Total natural marital fertility	Constant
TP_r	Total population Size in region r	Endogenous
$\alpha_{\mathrm{C}r}$	Regression intercept for $C_{\mathrm{C}r}$	Constant
$\alpha_{\mathrm{M}r}$	Regression intercept for $C_{\mathrm{M}r}$	Constant
$\beta_{\mathrm{EC}r}$	Education services regression coefficient for $C_{\mathrm{C}r}$	Constant
$\beta_{\mathrm{EM}r}$	Education services regression coefficient for $C_{\mathrm{M}r}$	Constant
$\beta_{\mathrm{HC}r}$	Health services regression coefficient for $C_{\mathrm{C}r}$	Constant
$\omega_{\mathrm{F}ijkr}$	Food shortage effect on death rate for age group i, sex j, and socioeconomic group k in region r	Constant
$\omega_{\mathrm{H}ijkr}$	Barriers to health services effect on death rate for age group i, sex j, and socioeconomic group k in region r	Constant

C.2 Health and Education Submodel Equatio and Variables

Sets

$$\text{Sex}: \quad \mathscr{S} = \{\text{Male}, \text{Female}\}$$

$$\text{Age Group}: \quad \mathscr{A} = \{\text{Age Group } 1, \ldots, \text{Age Group } 4\}$$

Equations

$$\text{Change in health services}: \frac{\mathrm{d}HS_r(t)}{\mathrm{d}t} = HG_r(t) - HD_r(t) \tag{C.15}$$

$$\text{Change in education services}: \frac{\mathrm{d}ES_r(t)}{\mathrm{d}t} = EG_r(t) - ED_r(t) \tag{C.16}$$

$$\text{Growth in health services}: HG_r(t) = \Lambda_{\mathrm{H}r} \cdot \Delta Y_r(t) \tag{C.17}$$

$$\text{Growth in education services}: EG_r(t) = \Lambda_{\mathrm{E}r} \cdot \Delta Y_r(t) \tag{C.18}$$

$$\text{Health service depreciation}: HD_r(t) = \zeta_{\mathrm{H}r} \cdot HS_r(t) \tag{C.19}$$

$$\text{Education service depreciation}: ED_r(t) = \zeta_{\mathrm{C}r} \cdot ES_r(t) \tag{C.20}$$

$$\text{Female health access}: HF_{kr}(t) = \chi_{\mathrm{HF1}kr} + \chi_{\mathrm{HF2}kr} \cdot \log\big(HS_r(t)\big) \tag{C.21}$$

$$\text{Female education attainment}: EF_{kr}(t) = \chi_{\mathrm{EF1}kr} + \chi_{\mathrm{EF2}kr} \cdot \log\big(ES_r(t)\big) \tag{C.22}$$

$$\begin{aligned}\text{General healthcare access}: HA_{kr}(t) &= \chi_{\mathrm{HA1}kr} + \chi_{\mathrm{HA2}kr} \cdot \log\big(HS_r(t)\big) \\ &\quad + \chi_{\mathrm{HA3}kr} \cdot \widehat{Y}_{kr}(t)\end{aligned} \tag{C.23}$$

Table C.2: Variables for Health and Education Submodel

Variables	Description	Type
ED_r	Depreciation education services in region r	Endogenous
EF_{kr}	Female education attainment for socioeconomic group k in region r	Endogenous
EG_r	Growth in education services in region r	Endogenous
ES_r	Education services availability in region r	Endogenous
HA_{kr}	General healthcare access for socioeconomic group k in region r	Endogenous
HD_r	Depreciation of health services in region r	Endogenous
HG_r	Growth in health services in region r	Endogenous
EF_{kr}	Female education attainment for socioeconomic group k in region r	Endogenous
HF_{kr}	Female access to maternal health services for socioeconomic group k in region r	Endogenous
HS_r	Health Services in region r	Endogenous
Y_r	Economic output in region r	Endogenous
$\widehat{Y}_{kr}$	Economic output per capita of socioeconomic group k in region r	Endogenous
$\Lambda_{\mathrm{E}r}$	Education development relative to econonomic output in region r	Exogenous
$\Lambda_{\mathrm{H}r}$	Health development relative to econonomic output in region r	Exogenous
$\chi_{\mathrm{EF1}kr}$	Fixed effect on female education attainment for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{HA1}kr}$	Fixed effect on healthcare access for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{HF1}kr}$	Fixed effect on female healthcare impact for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{EF2}kr}$	Education services impact on female education attainment for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{HA2}kr}$	Health services impact on healthcare access for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{HF2}kr}$	Health services impact female healthcare impact for socioeconomic group k in region r	Exogenous
$\chi_{\mathrm{HA3}kr}$	Income per capita impact on healthcare access for socioeconomic group k in region r	Exogenous
$\zeta_{\mathrm{H}r}$	Depreciation rate of health services in region r	Exogenous
$\zeta_{\mathrm{E}r}$	Depreciation rate of education services in region r	Exogenous

C.3 Economy Submodel Equations and Variables

Sets

Sex : $\mathscr{S} = \{\text{Male}, \text{Female}\}$

Age Group : $\mathscr{A} = \{\text{Age Group } 1, \ldots, \text{Age Group } 4\}$

Socioeconomic Group : $\mathscr{K} = \{\text{Rich}, \text{Poor}\}$

Equations

$$\text{Accessible Nonrenewable Resource}: N_r(t) = \Lambda_{\mathrm{N}r} \cdot N(t) \tag{C.24}$$

$$\text{Accessible Renewable Resource}: R_r(t) = \Lambda_{\mathrm{R}r} \cdot R(t) \tag{C.25}$$

$$\text{Economic output}: Y_r(t) = A_r \cdot L_r(t)^{\eta_{1r}} \cdot K_r(t)^{\eta_{2r}} \tag{C.26}$$

$$\text{Change in economic output}: \Delta Y_r(t) = Y_r(t) - Y_r(t-1) \tag{C.27}$$

$$\text{Income per capita}: \widehat{Y}_r(t) = \frac{Y_r(t)}{TP_r(t)} \tag{C.28}$$

$$\text{Change in income per capita}: \Delta\widehat{Y}_r(t) = \widehat{Y}_r(t) - \widehat{Y}_r(t-1) \tag{C.29}$$

$$\text{Change in capital}: \frac{\mathrm{d}K_r(t)}{\mathrm{d}t} = \Delta K_r(t) = I_r(t) - KD_r(t) \tag{C.30}$$

$$\text{Capital investment}: I_r(t) = \theta_{\mathrm{K}r} \cdot Y_r(t) \cdot N_r(t) \cdot R_r(t) \tag{C.31}$$

$$\text{Capital depreciation}: KD_r(t) = \iota_{\mathrm{K}} \cdot K_r(t) \tag{C.32}$$

$$\text{Employed labor force}: L_r(t) = \sum_{i \in \mathscr{A}} \sum_{j \in \mathscr{S}} \sum_{k \in \mathscr{K}} ER_{ijkr} \cdot P_{ijkr}(t) \tag{C.33}$$

$$\text{Inequality}: Q_r(t) = \Psi_r \cdot \frac{\Delta K_r(t)}{\Delta Y_r(t)} \tag{C.34}$$

APPENDIX C. EQUATIONS FOR A MULTI-COMPONENT GLOBAL POPULATION MODEL

Table C.3: Variables for Economic Submodel

Variables	Description	Type
A_r	Technology efficiency multiplier in region r	Constant
I_r	Capital investment flows in region r	Endogenous
K_r	Capital in region r	Endogenous
KD_r	Capital depreciation in region r	Endogenous
L_r	Employed labor force in region r	Endogenous
N	Global nonrenewable resources reserves	Endogenous
N_r	Nonrenewable resources reserves in region r	Endogenous
P_{ijkr}	Population of age i, sex j, and socioeconomic group k in region r	Endogenous
Q_r	Inequality in region r	Endogenous
ER_{ijkr}	Employment-to-working population ratio for age i, sex j, and socioeconomic group k in region r	Constant
R	Global renewable resource reserves	Endogenous
R_r	Renewable resource reserves in region r	Endogenous
TP_r	Total population in region r	Endogenous
Y_r	Economic output in region$r$	Endogenous
$\widehat{Y}_r$	Income per capita in region r	Endogenous
$\Delta\widehat{Y}_r$	Change in income per capita in region r	Endogenous
ι_{K}	Capital depreciation rate	Constant
$\Lambda_{\mathrm{N}r}$	Fraction of global nonrenewable resource reserves in region r	Constant
$\Lambda_(\mathrm{R}r)$	Fraction of global renewable resource reserves in region r	Constant
Ψ_r	Inequality parameter in region r	Constant
θ_{R}	Capital investment intensity in region r	Constant
$\eta_1 r$	Human capital production input elasticity in region r	Constant
$\eta_2 r$	Capital production input elasticity in region r	Constant

C.4 Global Natural Resources Submodel Equations and Variables

Sets

$$\text{Income Regions} : \mathscr{R} = \{\text{Low Income Countries},\ \text{Middle Income Countries},\ \text{High Income Countries}\}$$

Equations

$$\text{Change in renewable resource} : \frac{\mathrm{d}R(t)}{\mathrm{d}t} = RP(t) - RC(t) \tag{C.35}$$

$$\text{Replenishment of renewable resources} : RP = \delta_{\mathrm{R}} \cdot R(t) \tag{C.36}$$

$$\text{Change in nonrenewable resource} : \frac{\mathrm{d}N(t)}{\mathrm{d}t} = -NC(t) \tag{C.37}$$

$$\text{Renewable resources consumption} : RC(t) = \sum_{r \in \mathscr{R}} \iota_{\mathrm{R}r} \cdot Y_r(t) \tag{C.38}$$

$$\text{Nonrenewable resources consumption} : NC(t) = \sum_{r \in \mathscr{R}} \iota_{\mathrm{N}r} \cdot Y_r(t) \tag{C.39}$$

Table C.4: Variables for Global Natural Resources Submodel

Variables	Description	Type
N	Global nonrenewable resource reserves	Endogenous
NC	Nonrenewable resource consumption	Endogenous
R	Global renewable resource reserves	Endogenous
RC	Renewable resource consumption	Endogenous
RP	Renewable resource replenishment flow	Endogenous
Y_r	Economic output in region r	Endogenous
δ_{R}	Renewable resource replenishment rate	Constant
$\iota_{\mathrm{N}r}$	Nonrenewable resource consumption intensity in region r	Constant
$\iota_{\mathrm{R}r}$	Renewable resource consumption intensity in region r	Constant

C.5 Global Climate Submodel Equations and Variables

Sets

$$\text{Income Regions} : \mathscr{R} = \{\text{Low Income Countries},\ \text{Middle Income Countries},\ \text{High Income Countries}\}$$

Equations

Change in temperature :
$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = \lambda \cdot RF(t) \tag{C.40}$$

Radiative forcing :
$$RF(t) = 5.35 \cdot \ln \frac{G(t)}{G_0} + RF_{\mathrm{EXT}} \tag{C.41}$$

Change in CO_2 concentration :
$$\frac{\mathrm{d}G(t)}{\mathrm{d}t} = \widehat{G}(t) \cdot TP(t) - GS \tag{C.42}$$

CO_2 storage
$$GS = \Gamma \cdot G(t) \tag{C.43}$$

CO_2 emission :
$$G(t) = \sum_{r \in \mathscr{R}} \widehat{G}_r \cdot TP_r \tag{C.44}$$

CO_2 emission per capita :
$$\widehat{G}_r = \widehat{Y}_r(t)^{\psi_{\mathrm{E}}} \tag{C.45}$$

Table C.5: Variables for Climate Submodel

Variables	Description	Type
G	CO_2 concentration in the atmosphere	Endogenous
$\widehat{G}$	CO_2 emission per capita	Endogenous
G_0	Reference CO_2 concentration in the atmosphere	Constant
GS	CO_2 storage	Endogenous
RF	Radiative forcing from CO_2	Endogenous
RF_{EXT}	Radiative forcing from other GHG	Constant
T	Global temperature	Endogenous
TP_r	Total population in region r	Endogenous
$\widehat{Y}_r$	Income per capita in region r	Endogenous
Γ	Storage rate of CO_2	Constant
λ	Climate sensitivity	Constant
ψ_{E}	CO_2 emission rate relative to economic income	Constant

C.6 Global Water Resources Submodel Equations and Variables

Sets

$$\text{Income Regions } :\mathscr{R} = \{\text{Low Income Countries},\ \text{Middle Income Countries},\ \text{High Income Countries}\}$$

Equations

$$\textbf{Freshwater stock}: \frac{\mathrm{d}W(t)}{\mathrm{d}t} = WP(t) - WC(t) - WL(t) \tag{C.46}$$

$$\textbf{Water replenishment}: WP(t) = W(t) \cdot NWR \tag{C.47}$$

$$\textbf{Water loss due to climate change}: WL(t) = CC_{\mathrm{W}}(t) \cdot W(t) \tag{C.48}$$

$$\textbf{Water loss rate due to climate}: CC_{\mathrm{W}}(t) = \delta_{\mathrm{W}} \cdot \frac{T_0}{T(t)} \tag{C.49}$$

$$\textbf{Municipal demand for water}: WD_{\mathrm{M}}(t) = \sum_{r \in \mathscr{R}} \widehat{WR}_{\mathrm{M}r} \cdot TP_r(t) \tag{C.50}$$

$$\textbf{Industrial demand for water}: WD_{\mathrm{I}}(t) = \zeta_{\mathrm{I}} \cdot \sum_{r \in \mathscr{R}} Y_r(t) \tag{C.51}$$

$$\textbf{Global demand for water}: WD(t) = WD_{\mathrm{M}}(t) + WD_{\mathrm{A}}(t) + WD_{\mathrm{I}}(t) \tag{C.52}$$

$$\textbf{Water consumption}: WC(t) = \min\big\{WD(t), \tag{C.53}$$

$$\big(1 - CC_{\mathrm{W}}(t)\big) \cdot W(t)\big\} \tag{C.54}$$

APPENDIX C. EQUATIONS FOR A MULTI-COMPONENT GLOBAL POPULATION MODEL

Table C.6: Variables for Water Submodel

Variables	Description	Type
CC_{W}	Climate change effects on water supply	Endogenous
NWR	Natural water replenishment rate	Constant
P_{ij}	Population count in age group i and sex j	Endogenous
T	Global temperature	Endogenous
T_0	Initial global temperature	Constant
W	Freshwater supply (Total renewable actual water sources)	Endogenous
WC	Global water consumption rate	Endogenous
WD	Global water demand	Endogenous
WD_{A}	Agricultural demand for water	Endogenous
$\widehat{WD}_{\mathrm{M}r}$	Municipal water demand per capita in region r	Endogenous
WD_{M}	Municipal demand for water	Endogenous
WD_{I}	Industrial demand for water	Endogenous
WL	Water loss due to climate change	Endogenous
WP	Water replenishment	Endogenous
Y_r	Economic Output in region r	Endogenous
δ_{W}	Water loss sensitivity to climate change	Constant
ζ_I	Consumption of water per unit of economic output Y	Constant

C.7 Food Submodel Equations and Variables

Sets

$$\text{Sex}: \mathscr{S} = \{\text{Male}, \text{Female}\}$$

$$\text{Age Group}: \mathscr{A} = \{\text{Age Group 1}, \ldots, \text{Age Group 4}\}$$

$$\text{Socioeconomic Group}: \mathscr{K} = \{\text{Rich}, \text{Poor}\}$$

$$\text{Income Regions}: \mathscr{R} = \{\text{Low Income Countries}, \text{Middle Income Countries}, \text{High Income Countries}\}$$

$$\text{Food Types}: \mathscr{F} = \{\text{Fish}, \text{Livestock}, \text{Crops}\}$$

Equations

$$\text{Change in Econ. Output per capita}: \Delta\widehat{Y}_r(t) = \widehat{Y}_r(t) - \widehat{Y}_r(t-1) \tag{C.55}$$

$$\text{Change in Food Demand per capita}: \frac{\mathrm{d}\widehat{FD}_r(t)}{\mathrm{d}t} = \epsilon_r \frac{\Delta\widehat{Y}_r(t) \cdot \widehat{FD}_r(t-1)}{\widehat{Y}_r(t-1)} \tag{C.56}$$

$$\text{Global Demand for All Food}: FD(t) = \widehat{FD}(t) \cdot \sum_{r \in \mathscr{R}} TP_r(t) \tag{C.57}$$

$$\text{Global Demand for Food } m: FD_m(t) = \sigma_m \cdot FD(t) \tag{C.58}$$

$$\text{Change in Global Stock of Food } m: \frac{\mathrm{d}FS_m(t)}{\mathrm{d}t} = FP_m(t) - FW_m(t) - FC_m(t) \tag{C.59}$$

$$\text{Food Type } m \text{ Waste}: FW_m(t) = \delta_{\mathrm{F}m} \cdot FS_m(t) \tag{C.60}$$

$$\text{Consumption of Food Type } m: FC_m(t) = \min\left\{FS_m(t),\ \gamma_m \cdot FD_m(t)\right\} \tag{C.61}$$

$$\text{Food Access for the Rich}: FA_{\mathrm{RICH}r}(t) = \Omega_r \cdot Q_r(t) \tag{C.62}$$

$$\text{Food Access for the Poor}: FA_{\mathrm{POOR}r}(t) = \left(1 - FA_{\mathrm{RICH}r}(t)\right) \tag{C.63}$$

$$\text{Nutritional Consumption}: NC_{kr}(t) = FA_{kr}(t) \tag{C.64}$$

$$\cdot \sum_{m \in \mathscr{F}} \delta_m \cdot FC_m(t) \tag{C.65}$$

$$\text{Nutritional Cons. per capita}: \widehat{NC}_{kr} = \frac{NC_k(t)}{\sum_{i\in\mathscr{A}}\sum_{j\in\mathscr{S}} P_{ijkr}(t)} \tag{C.66}$$

$$\text{Food } m \text{ Production}: FP_m(t) = \max\left\{\min\left\{\overline{FP}_m(t), \frac{\widetilde{FP}_m(t) - FS_m(t)}{\kappa_m}\right\}, 0\right\} \tag{C.67}$$

$$\text{Target Food Production}: \widetilde{FD}_m(t) = \gamma_m \cdot FD(t) \tag{C.68}$$

$$\text{Livestock and Crop Production Capacity}: \overline{FC}_m(t) = A_{\mathrm{F}m} \cdot \left[S_m(t)\right]^{\nu_{1m}} \cdot \left[\xi_{\mathrm{W}m} \cdot W(t)\right]^{\nu_{2m}} \tag{C.69}$$

$$\text{Fish Production Capacity}: \overline{FP}_{\mathrm{FISH}}(t) = \zeta_U \cdot U(t) \tag{C.70}$$

$$\text{Change in Fisheries Stock}: \frac{\mathrm{d}U(t)}{\mathrm{d}t} = UP(t) - FP_{\mathrm{FISH}}(t) \tag{C.71}$$

$$\text{Fisheries Replenishment}: UP(t) = \theta_{\mathrm{U}} \cdot U(t) \tag{C.72}$$

$$\text{Water Demand for Production of Food } m: WD_{\mathrm{A}m}(t) = \frac{1}{\xi_{\mathrm{W}m}} \cdot \left(\frac{\widehat{FD}(t)}{A_{\mathrm{F}m} \cdot \left[S_m(t)\right]^{\nu_{1m}}}\right)^{\frac{1}{\nu_{2m}}} \tag{C.73}$$

$$\text{Total Agricultural Water Demand}: WD_{\mathrm{A}}(t) = \sum_{m\in\mathscr{F}} WD_{\mathrm{A}m}(t) \tag{C.74}$$

$$\text{Change in Agricultural Land Stock}: \frac{\mathrm{d}S_m(t)}{\mathrm{d}t} = LP_m(t) - LL_m(t) \tag{C.75}$$

$$\text{Agricultural Land Gain} LP(t) = \theta_{\mathrm{L}m} \cdot S_m(t) \tag{C.76}$$

$$\text{Agricultural Land Loss} LL(t) = \delta_{\mathrm{L}m} \cdot S_m(t) \tag{C.77}$$

APPENDIX C. EQUATIONS FOR A MULTI-COMPONENT GLOBAL POPULATION MODEL

Table C.7: Variables for Food Submodel

Variables	Description	Type
$A_{\mathrm{F}m}$	Technology multiplier for food production type m	Constant
FC_m	Consumption of food type m	Endogenous
$\widehat{NC}_{ij}$	Nutrition consumption per capita by population group age i sex j	Endogenous
$\widehat{FA}_k$	Food access for socioeconomic group k	Endogenous
$\widehat{FD}_m$	Food demand per capita for food type m	Endogenous
$\Delta\widehat{FD}_m$	Change in food demand for food type m	Endogenous
FD	Global food demand for food type m	Endogenous
FD_m	Food demand for food type m	Endogenous
FP_m	Production of food type m	Endogenous
FP_{FISH}	Production of fish	Endogenous
$\overline{FP}$	Production capacity of food type m	Endogenous
$\overline{FP}_{\mathrm{FISH}}$	Production capacity of fish	Endogenous
$\widetilde{FP}_m$	Target production of food type m	Endogenous
FS_m	Food stock for food type m	Endogenous
FW_m	Food waste of food type m	Endogenous
LC_m	Agricultural land loss of food type m due to other factors	Endogenous
LL_m	Agricultural land loss of food type m due to climate change	Endogenous
LP_m	Agricultural conversion of land food type m	Endogenous
$\widehat{NR}$	Food requirement per capita	Constant
P_{ij}	Population group age i and sex j	Endogenous
Q_r	Inequality in region r	Endogenous
S_m	Agricultural land area for food type m	Endogenous
U	Fisheries stock	Endogenous
UP	Fisheries replenishment	Endogenous
W	Water stock	Endogenous
$\widetilde{WC}_m$	Desired consumption of water for food production of type m	Endogenous
$WD_{\mathrm{A}m}$	Agricultural water demand from food production of type m	Endogenous
$\widehat{Y}_r$	Income per capita in region r	Constant
$\Delta\widehat{Y}_r$	Change in income per capita in region r	Constant
δ_{L}	Fraction of agricultural land loss per year	Constant
$\delta_{\mathrm{F}m}$	Percentage of food type m that is wasted	Constant
ϵ_m	Income elasticity for demand of food type m	Constant
ζ_{U}	Fraction of fisheries that are accessible	Constant
γ_m	Nutritional conversion for food type m	Constant
κ_m	Speed of adjustment for price expectations of food type m	Constant
ν_{1m}	Production elasticity for land inputs of food type m	Constant
ν_{2m}	Production elasticity for water inputs of food type m	Constant
$\theta_{\mathrm{L}m}$	Fraction of agricultural land gained per year	Constant
θ_{U}	Fisheries replenishment rate	Constant
σ_m	Fraction of demand for food type m	Constant
$\xi_{\mathrm{W}m}$	Fraction of freshwater supply dedicated to food production m	Constant
Ω_{kr}	Inequality effect on food access for socioeconomic group k in region r	Constant

Appendix D

Calibration Plots for a Health and Education Services of Multi-component Global Population Model

APPENDIX D. CALIBRATION PLOTS FOR A HEALTH AND EDUCATION SERVICES OF MULTI-COMPONENT GLOBAL POPULATION MODEL

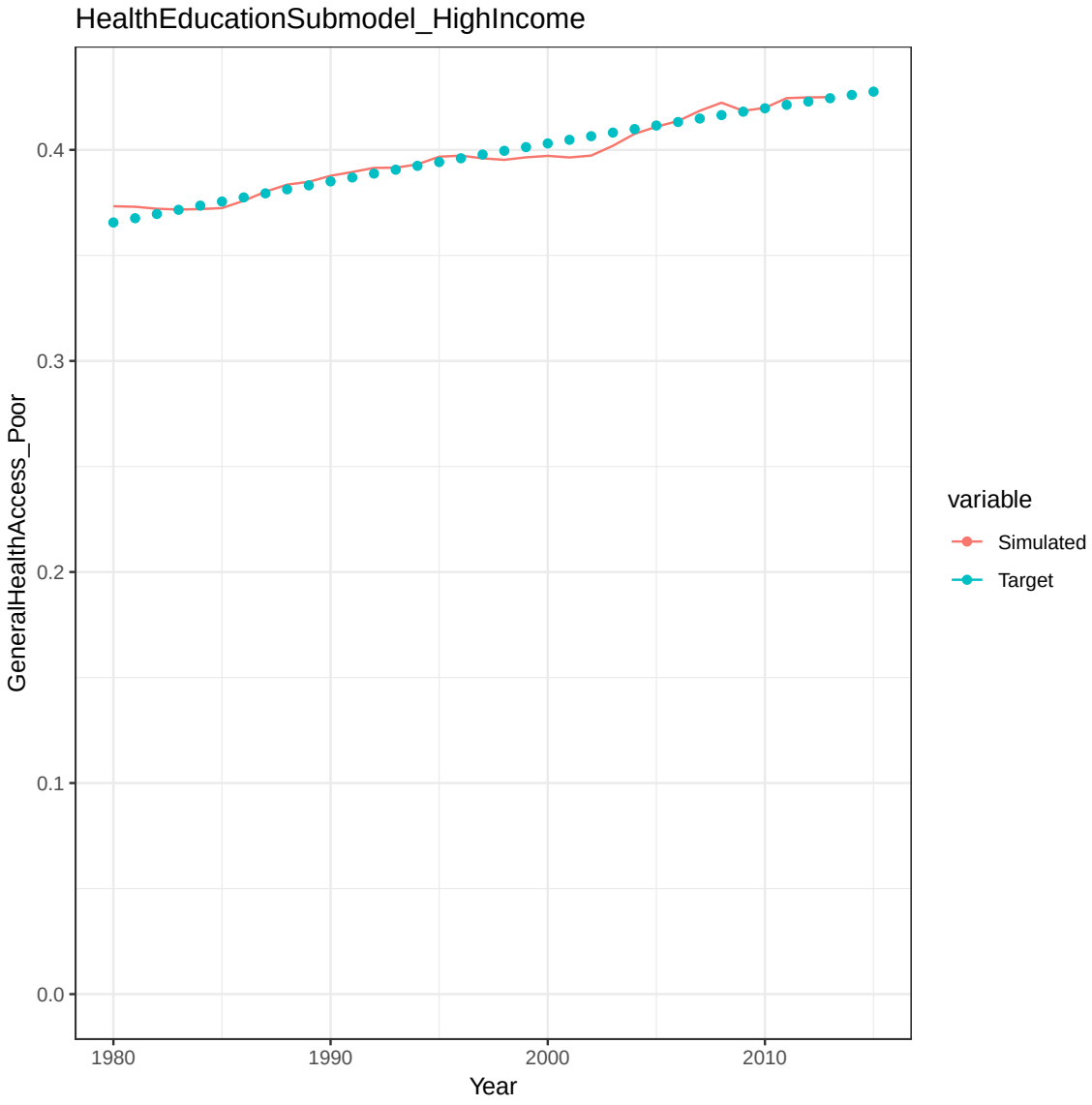


Figure D.1: General Health Access of Poor Population in High Income Region.

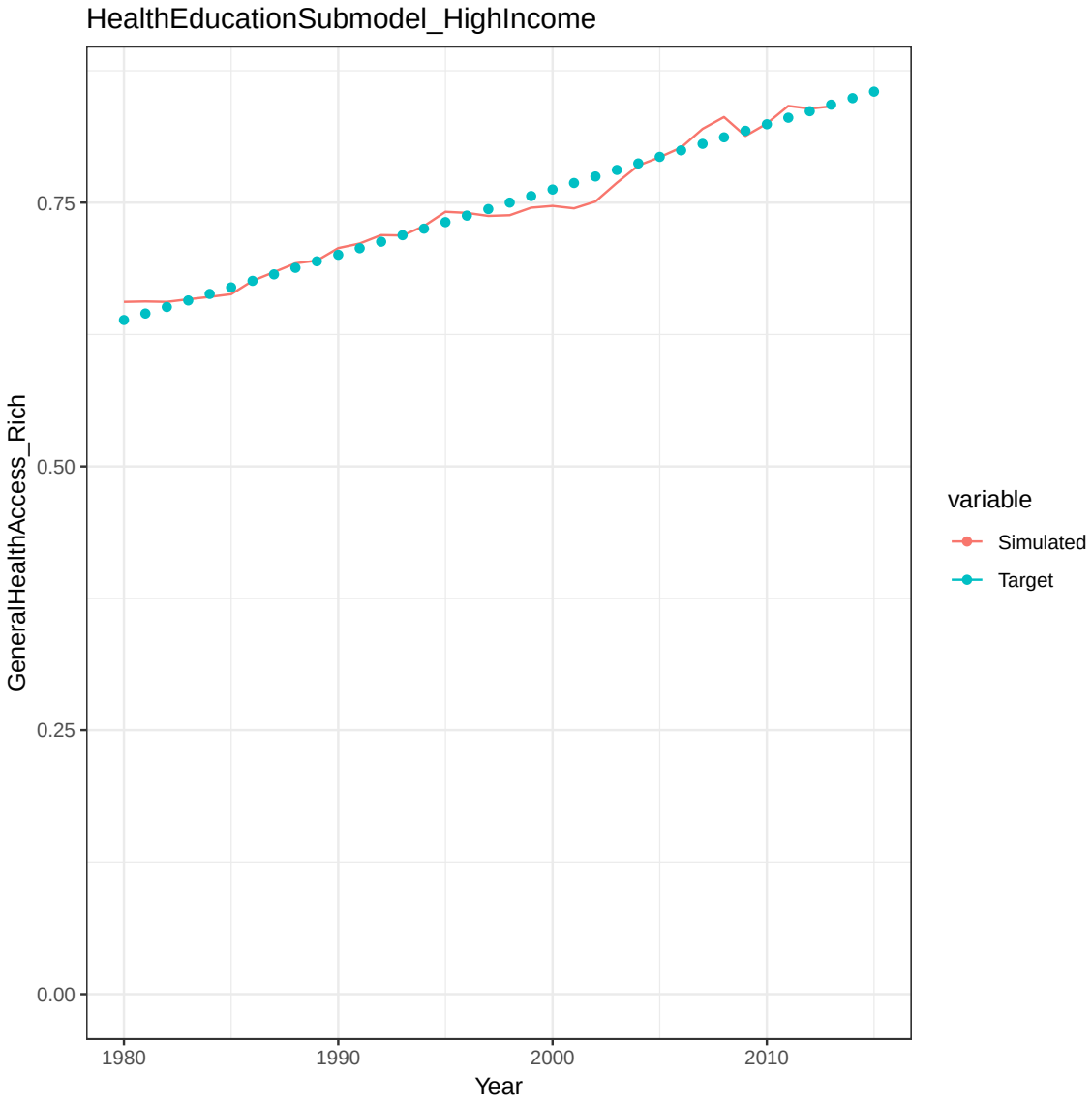


Figure D.2: General Health Access of Rich Population in High Income Region.

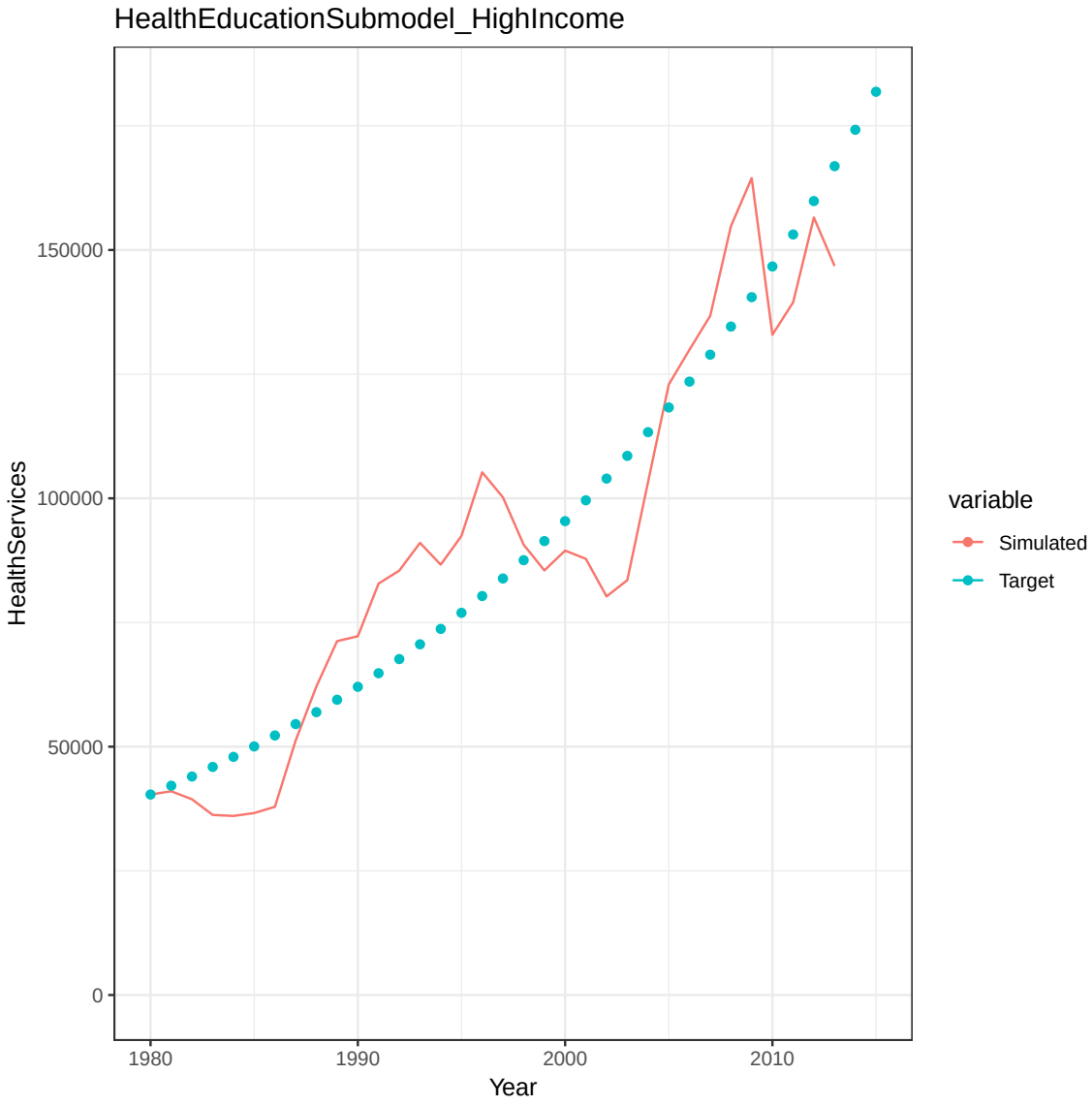


Figure D.3: Health Services in High Income Region.

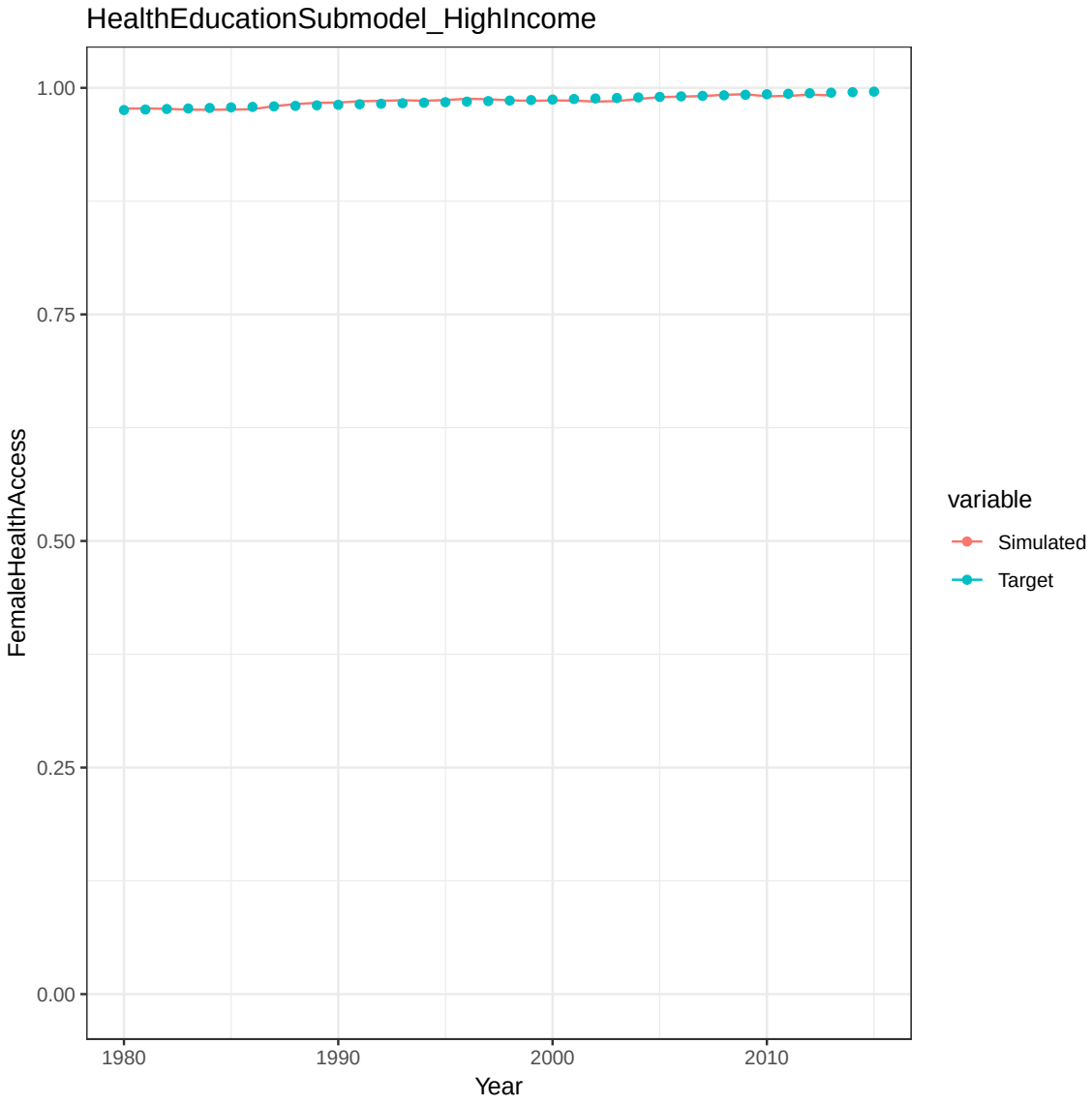


Figure D.4: Female Health Access in High Income Region.

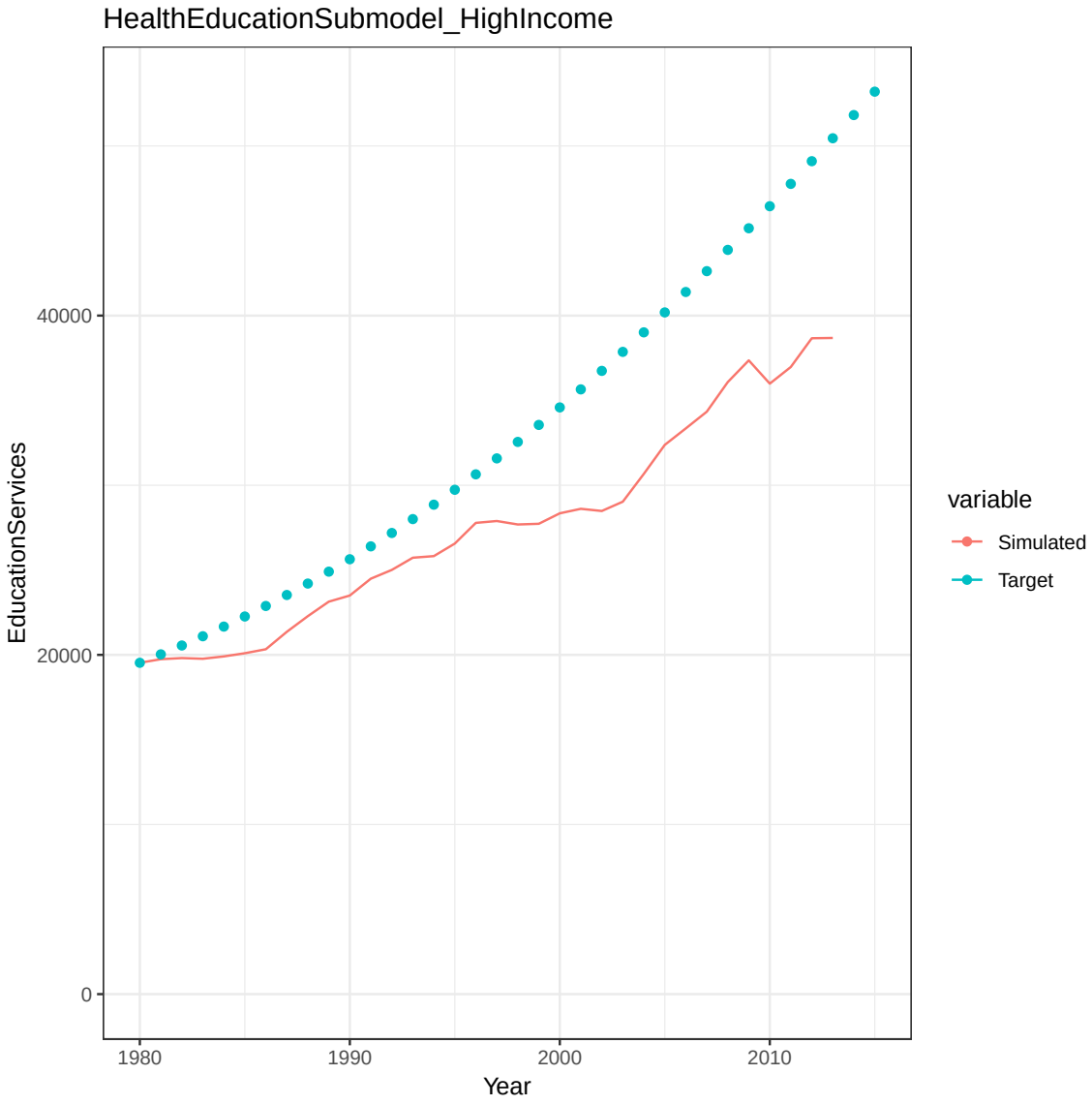


Figure D.5: Education Services in High Income Region.

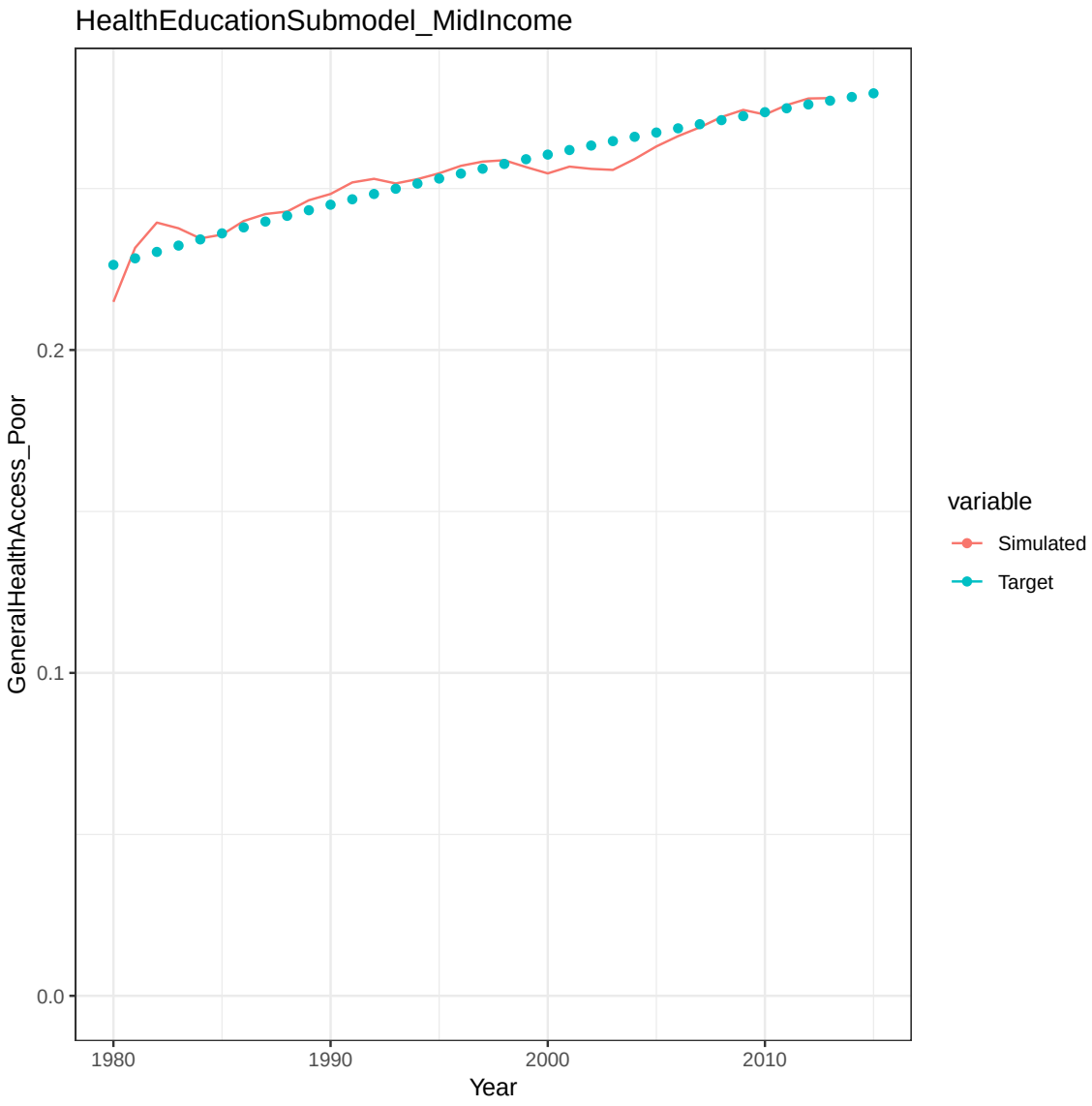


Figure D.6: General Health Access of Poor Population in Middle Income Region.

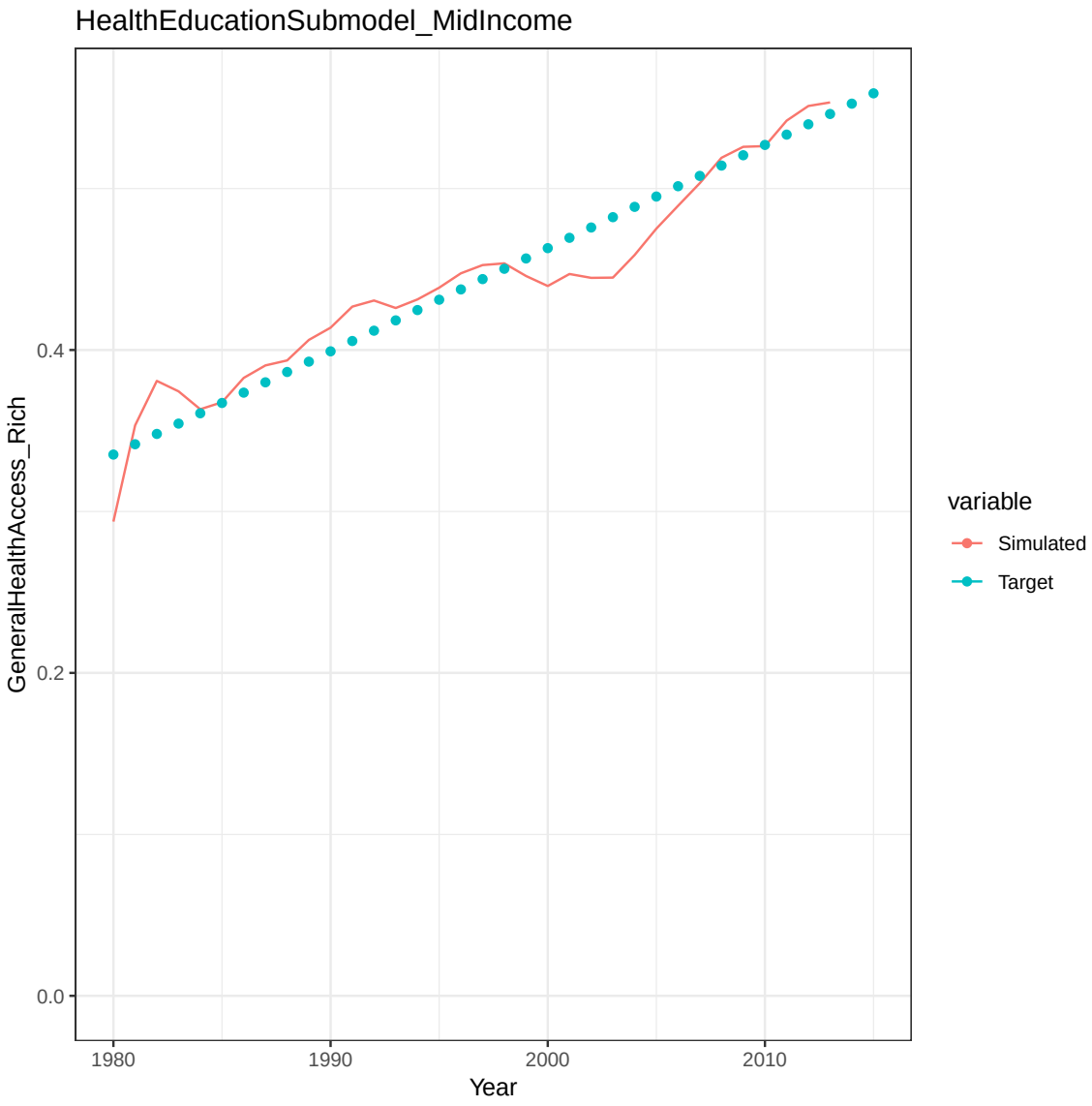


Figure D.7: General Health Access of Rich Population in Middle Income Region.

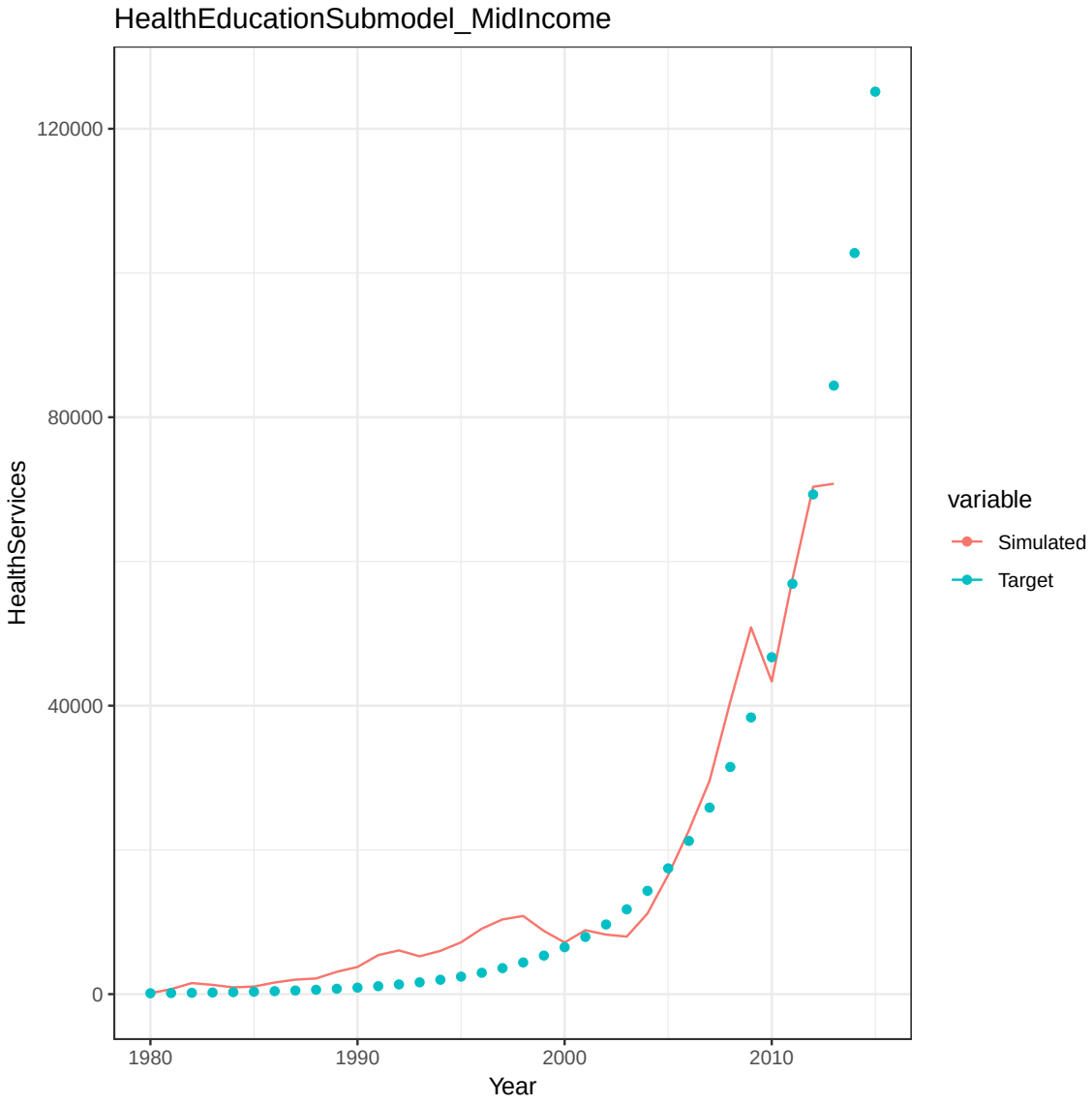


Figure D.8: Health Services in Middle Income Region.

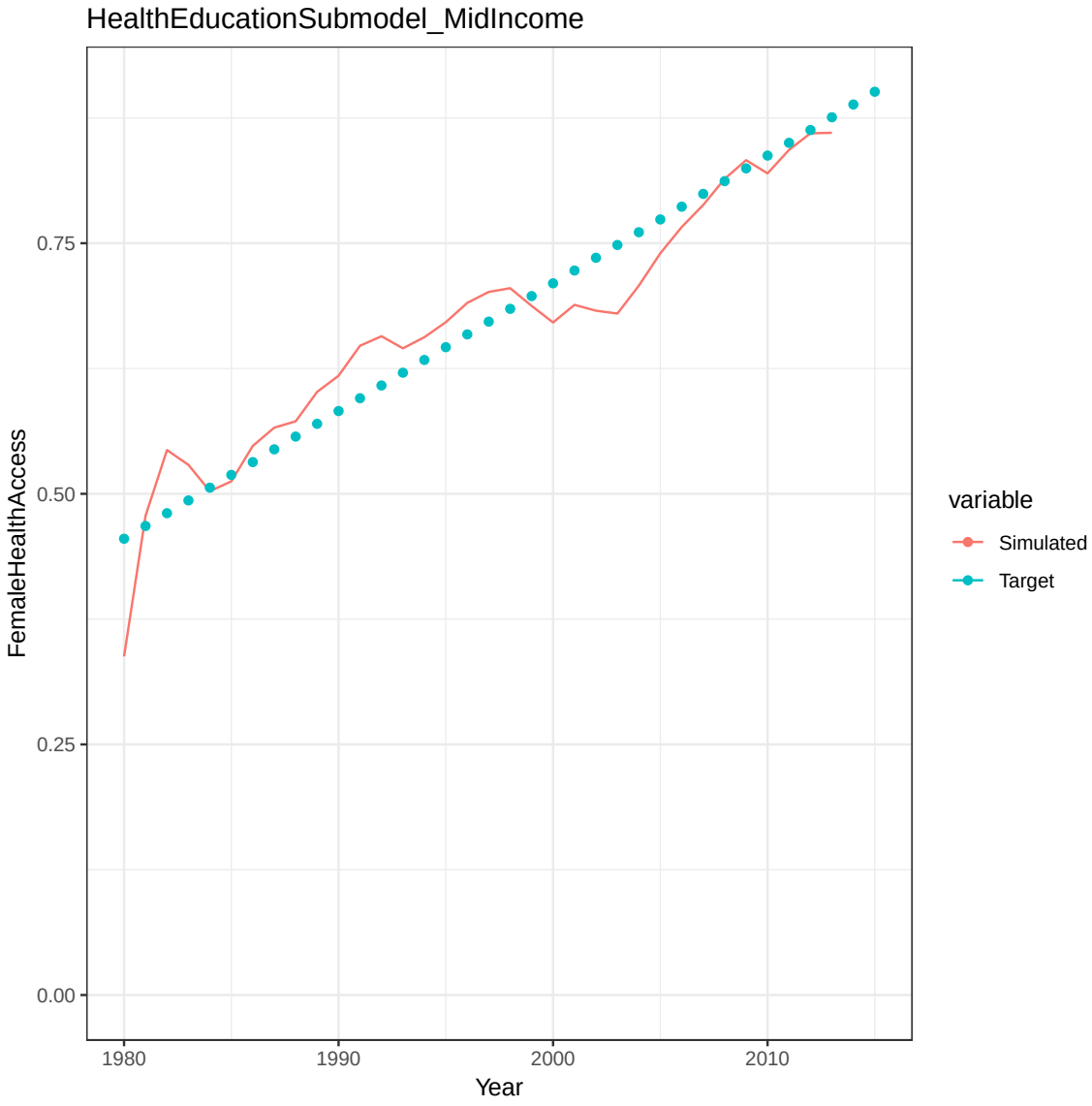


Figure D.9: Female Health Access in Middle Income Region.

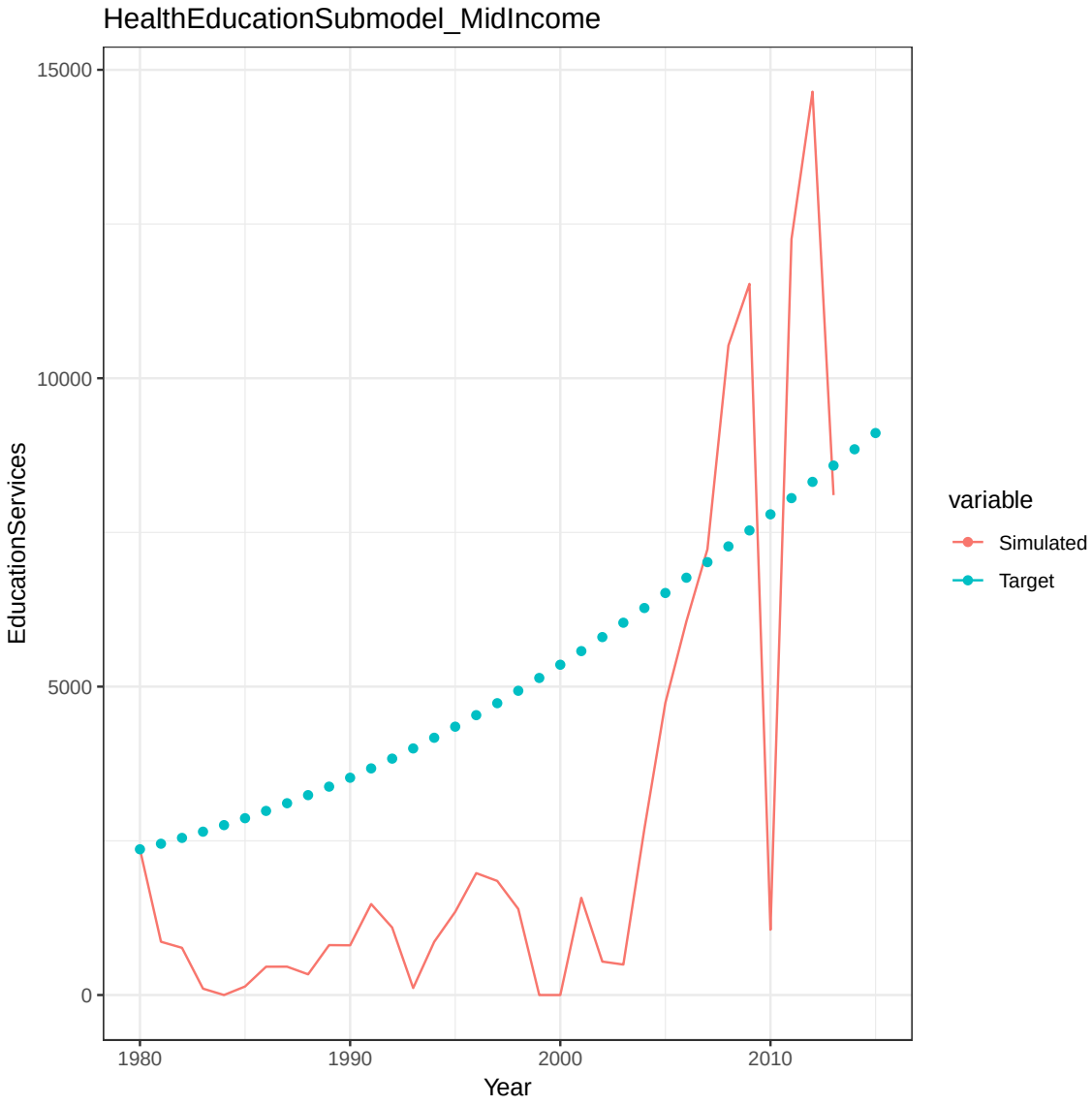


Figure D.10: Education Services in Middle Income Region.

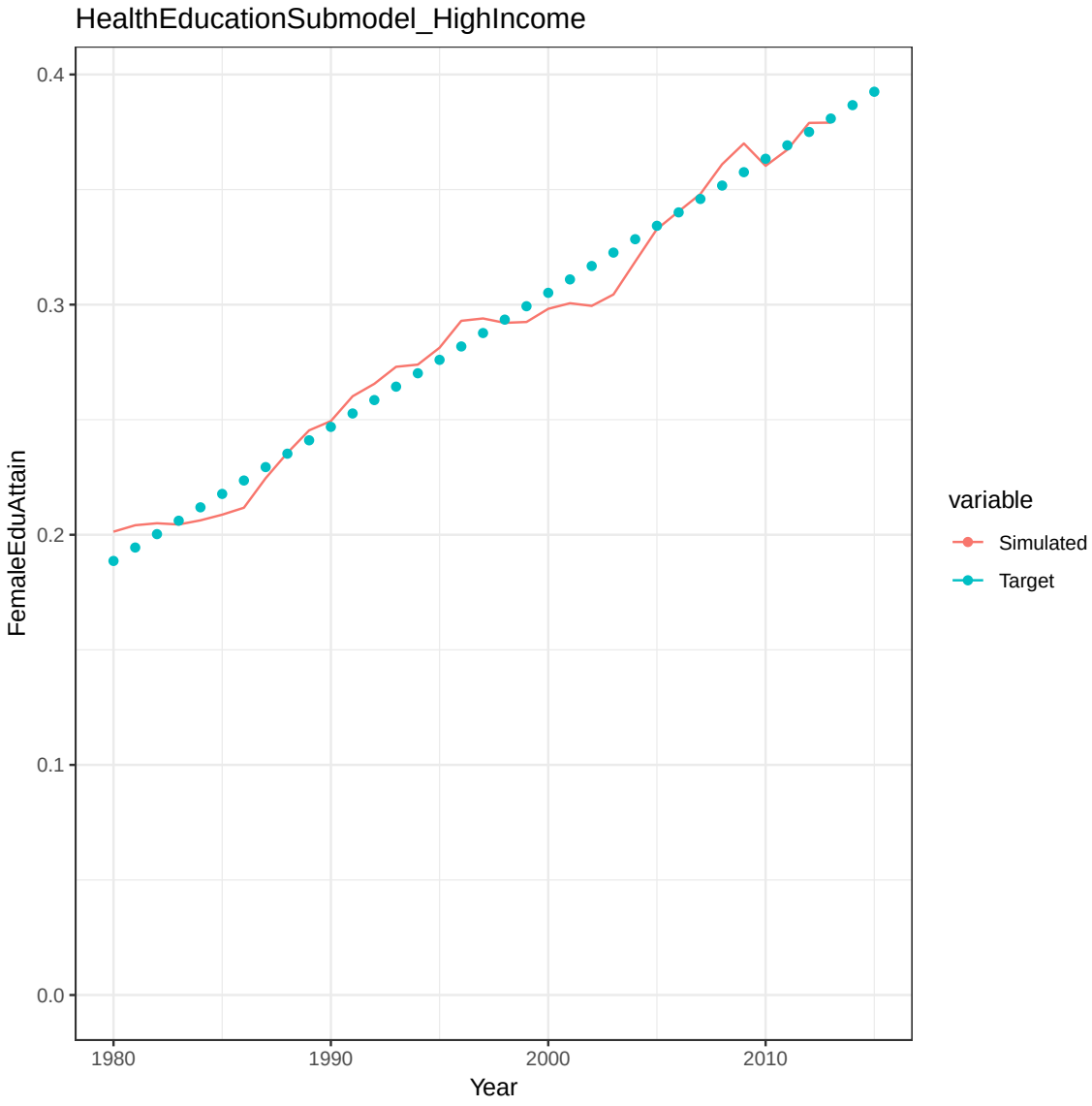


Figure D.11: Female Education Attainment in High Income Region.

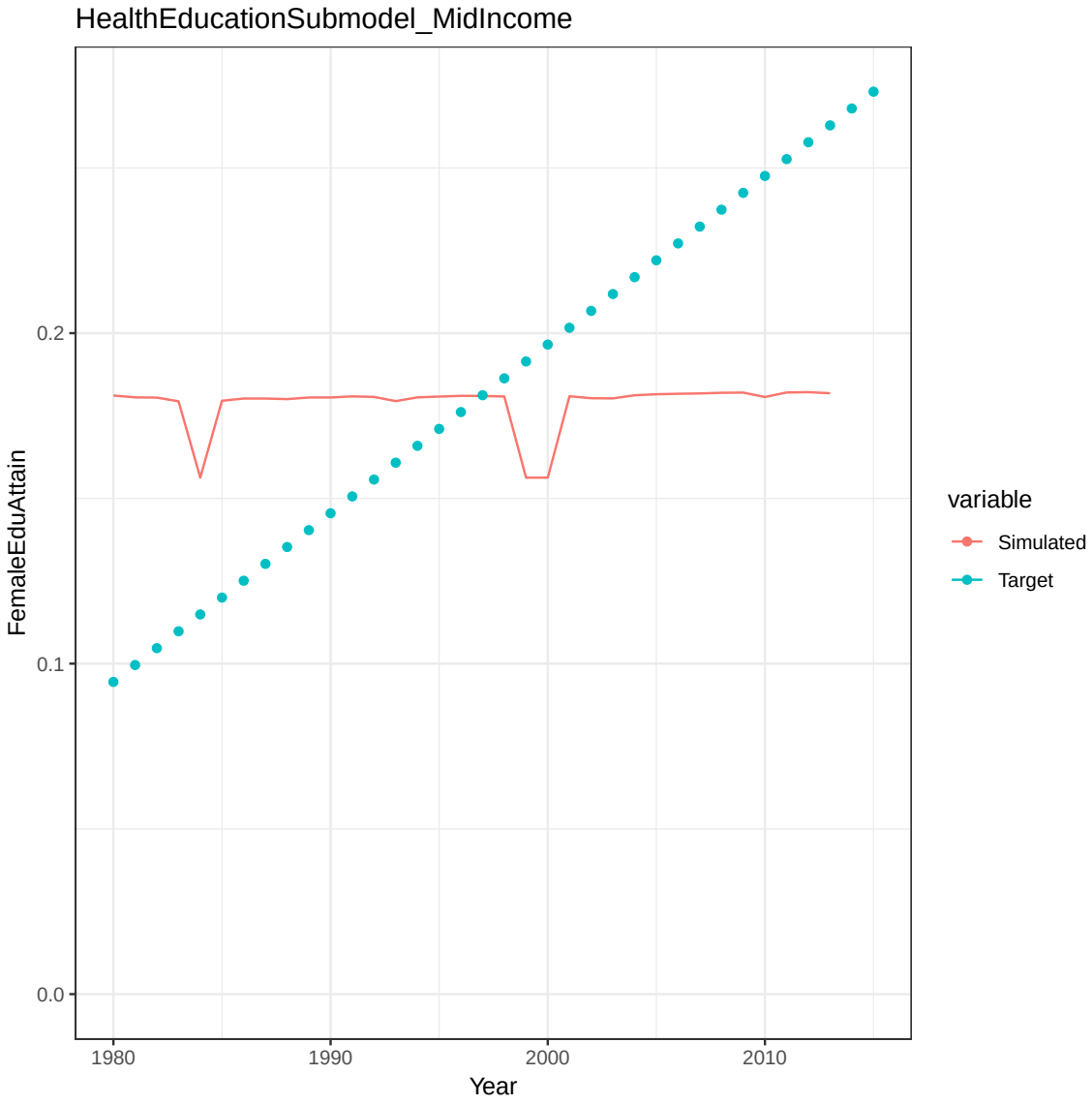


Figure D.12: Female Education Attainment in Middle Income Region.

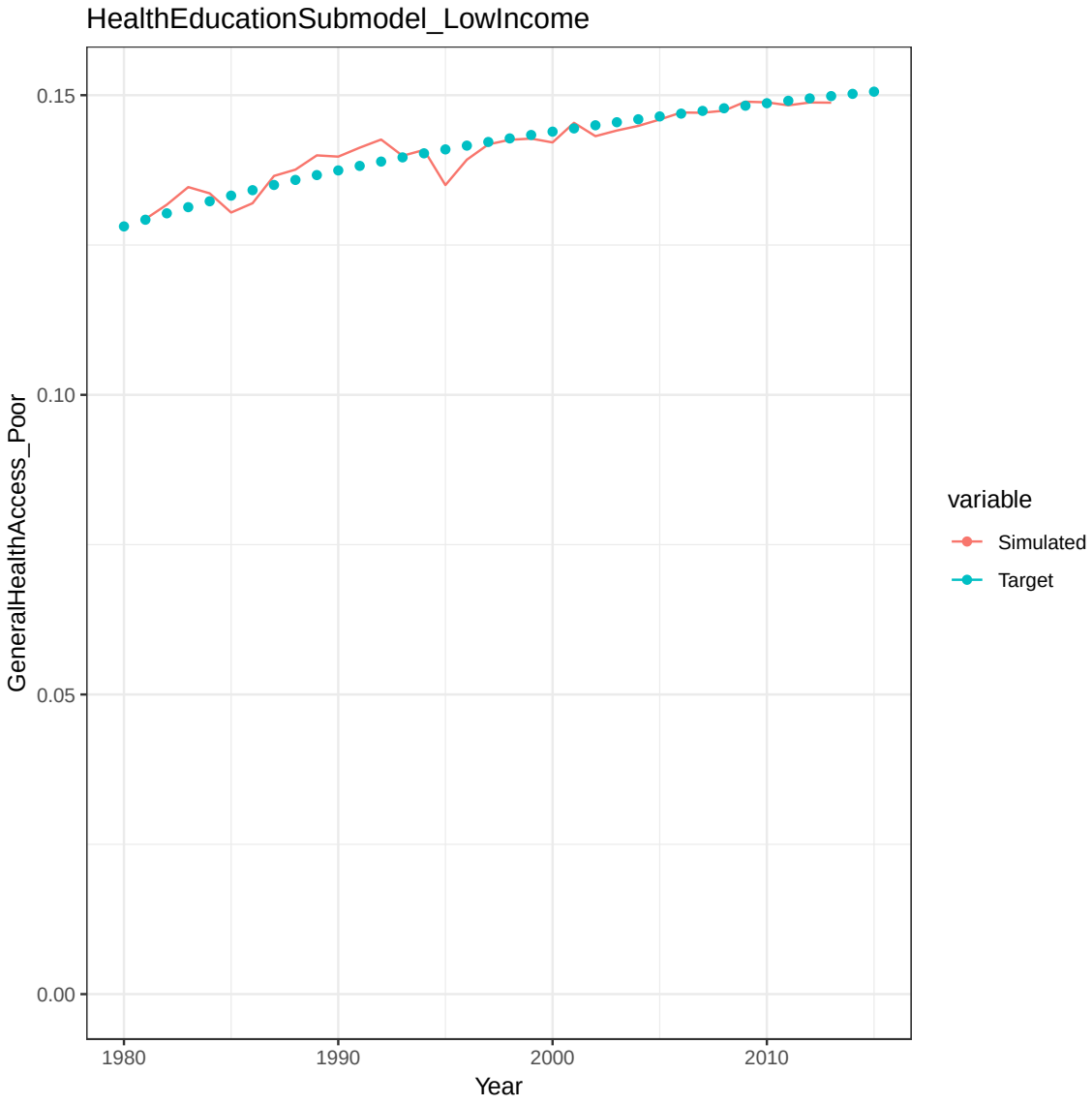


Figure D.13: General Health Access of Poor Population in Low Income Region.

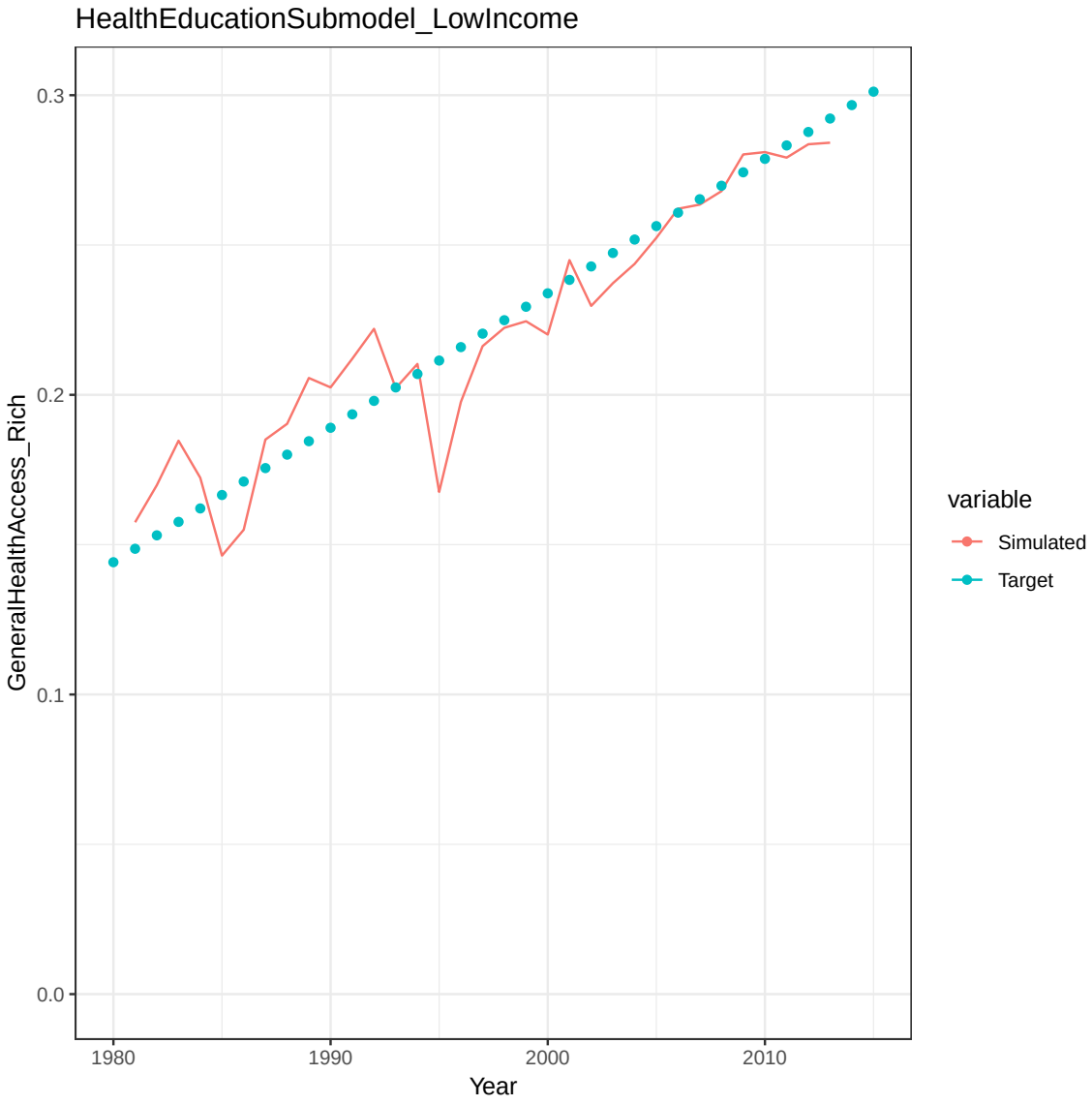


Figure D.14: General Health Access of Rich Population in Low Income Region.

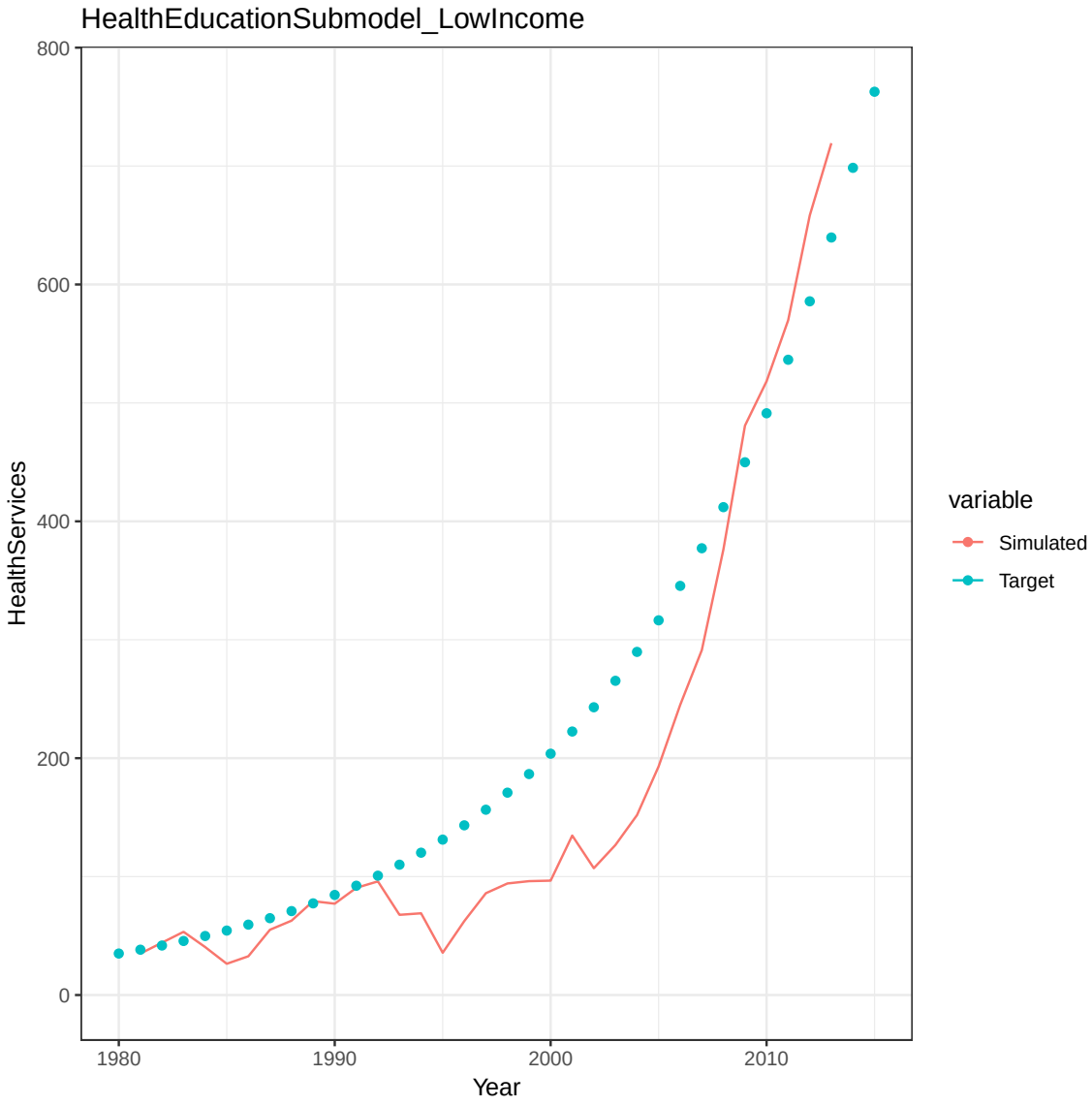


Figure D.15: Health Services in Low Income Region.

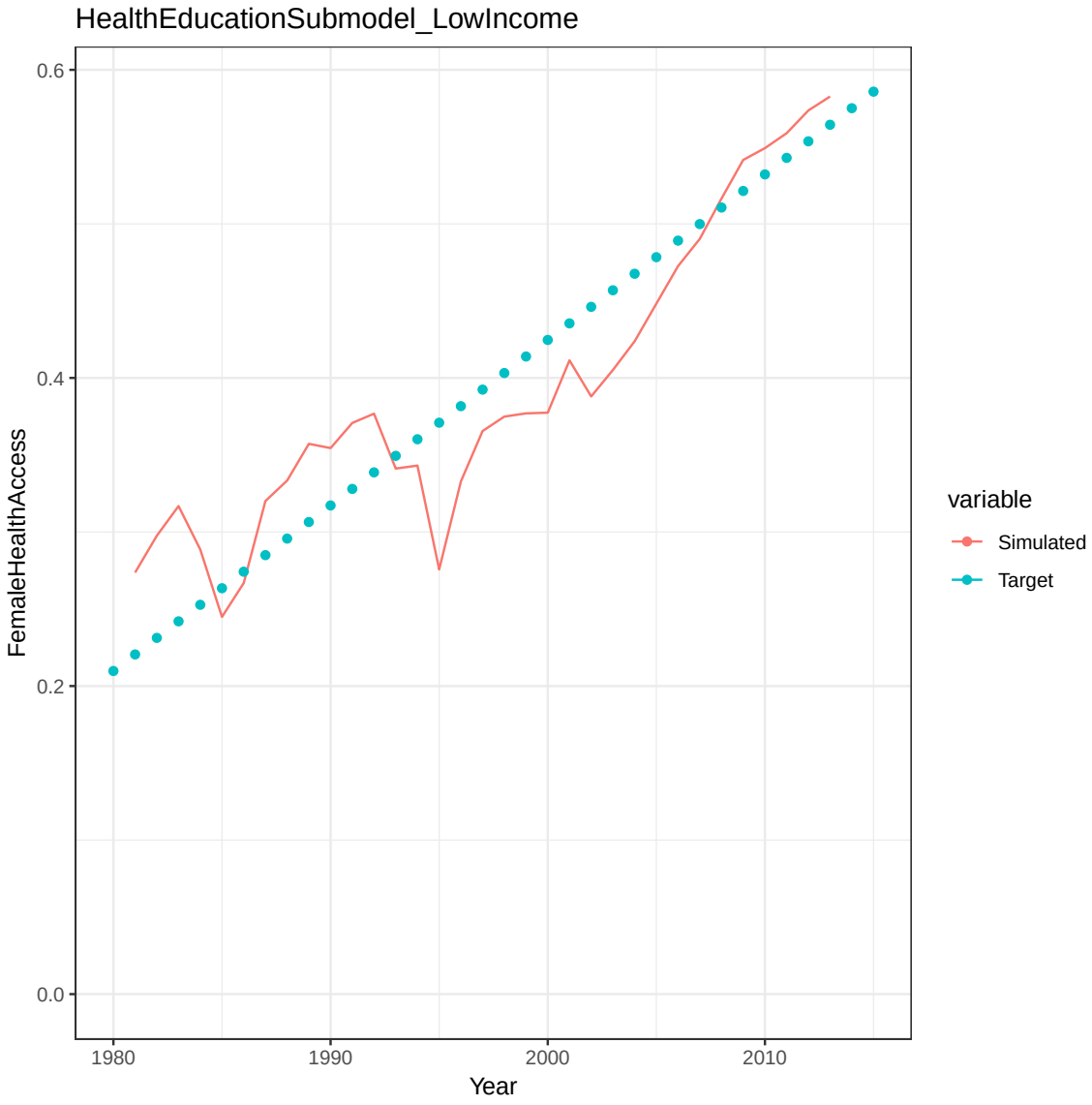


Figure D.16: Female Health Access in Low Income Region.

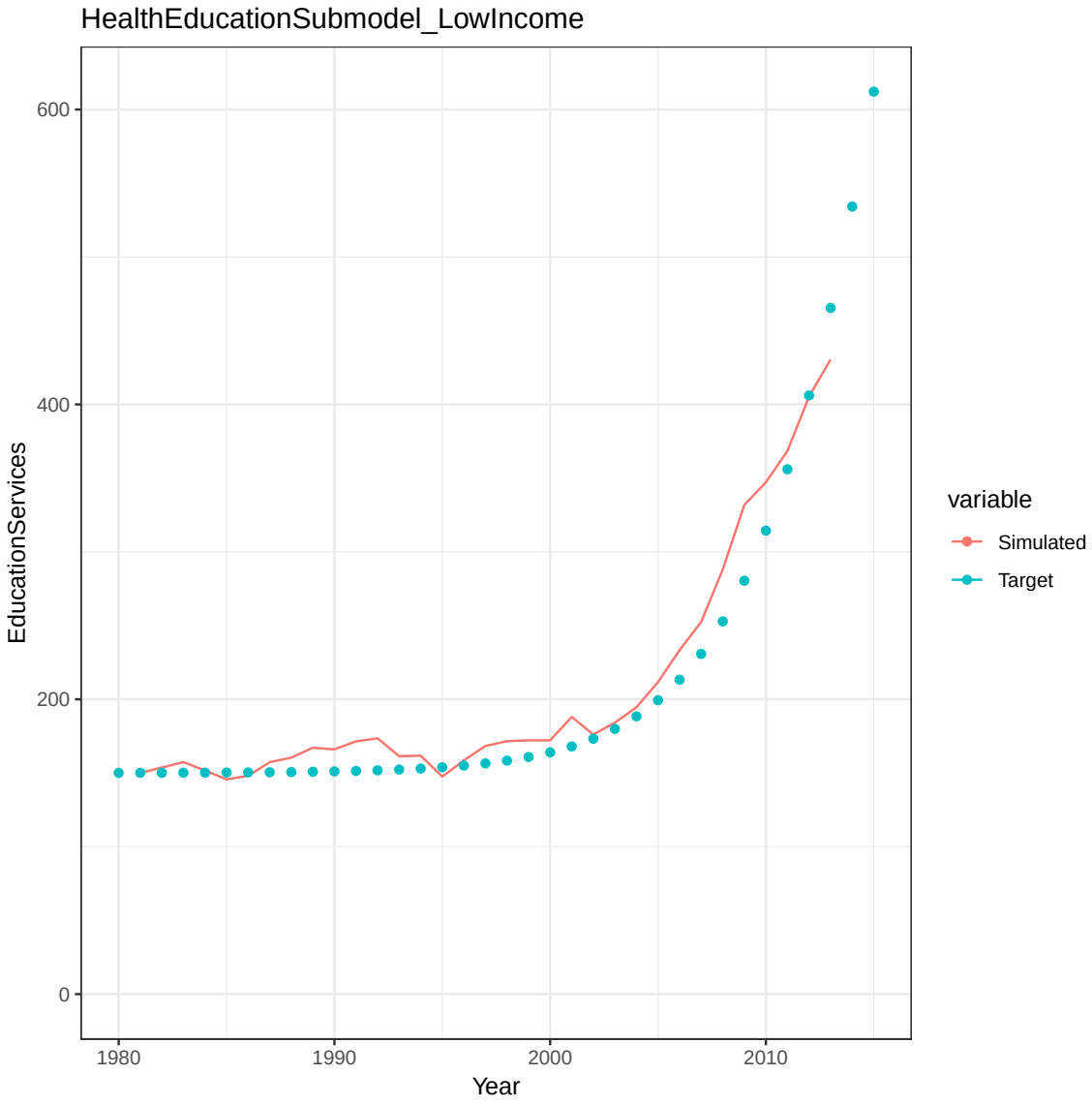


Figure D.17: Education Services in Low Income Region.

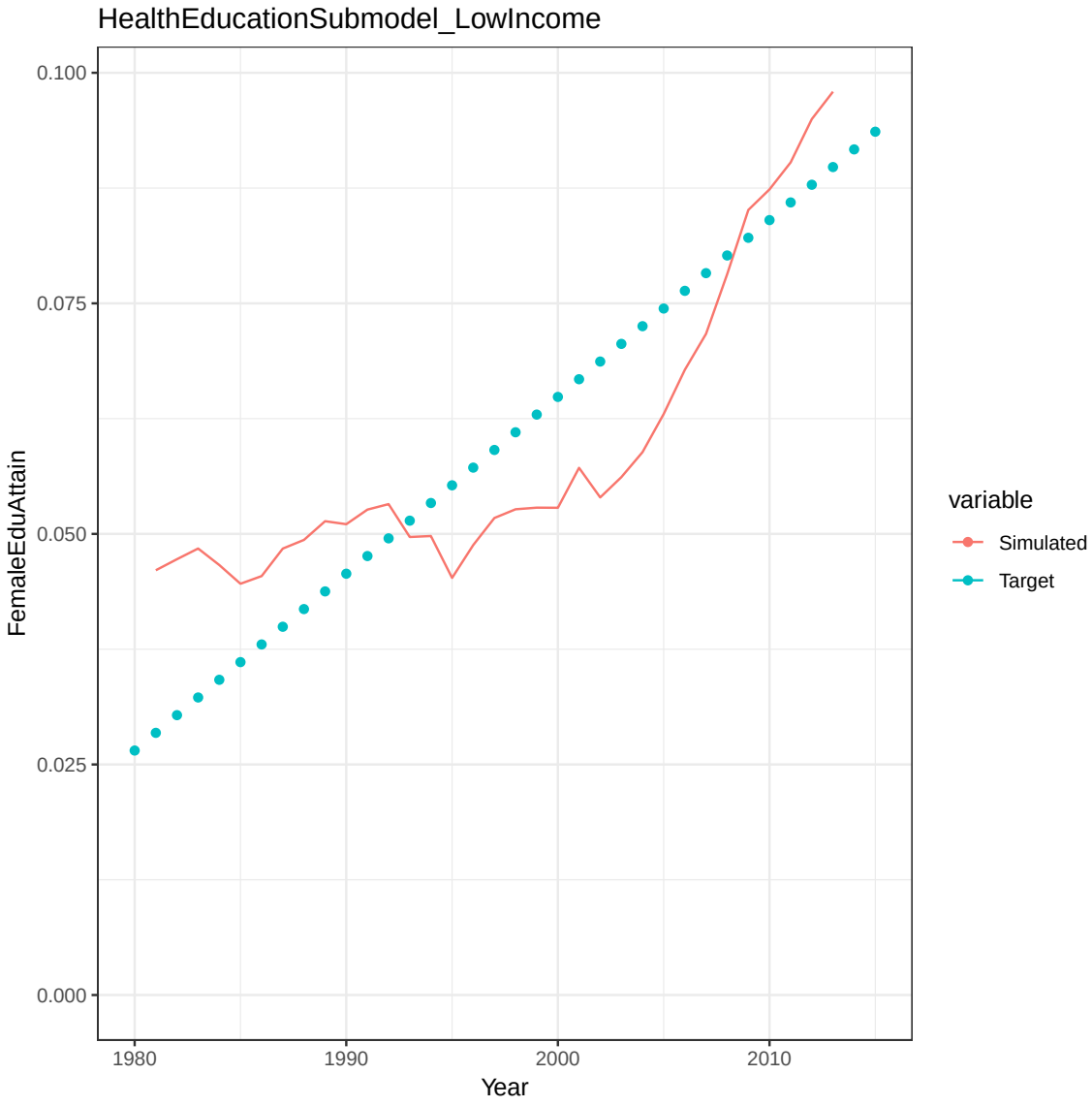


Figure D.18: Female Education Attainment in Low Income Region.

Bibliography

[1] Y. Bar-Yam, *Dynamics of complex systems*. Addison-Wesley Reading, MA, 1997, vol. 213.

[2] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[3] J. B. Homer and G. B. Hirsch, "System dynamics modeling for public health: background and opportunities," *American journal of public health*, vol. 96, no. 3, pp. 452–458, 2006.

[4] J. D. Sterman, "Learning from evidence in a complex world," *American journal of public health*, vol. 96, no. 3, pp. 505–514, 2006.

[5] B. Madahian, R. C. Klesges, L. Klesges, and R. Homayouni, "System dynamics modeling of childhood obesity," in *BMC bioinformatics*, vol. 13, no. 12. BioMed Central, 2012, p. A13.

BIBLIOGRAPHY

[6] R. A. Hammond, "Peer reviewed: complex systems modeling for obesity research," *Preventing chronic disease*, vol. 6, no. 3, 2009.

[7] S. Allender, B. Owen, J. Kuhlberg, J. Lowe, P. Nagorcka-Smith, J. Whelan, and C. Bell, "A community based systems diagram of obesity causes," *PLoS One*, vol. 10, no. 7, p. e0129683, 2015.

[8] H. Frumkin, J. Hess, G. Luber, J. Malilay, and M. McGeehin, "Climate change: the public health response," *American journal of public health*, vol. 98, no. 3, pp. 435–445, 2008.

[9] D. De Savigny and T. Adam, *Systems thinking for health systems strengthening*. World Health Organization, 2009.

[10] J. Mingers and L. White, "A review of the recent contribution of systems thinking to operational research and management science," *European Journal of Operational Research*, vol. 207, no. 3, pp. 1147–1161, 2010.

[11] J. D. Sterman, *Business dynamics: systems thinking and modeling for a complex world*, 2000, no. HD30. 2 S7835 2000.

[12] S. Friedenthal, A. Moore, and R. Steiner, *A practical guide to SysML: the systems modeling language*. Morgan Kaufmann, 2014.

[13] J. W. Forrester, "System dynamics, systems thinking, and soft or," *System dynamics review*, vol. 10, no. 2-3, pp. 245–256, 1994.

[14] ——, "Industrial dynamics," *Journal of the Operational Research Society*, vol. 48, no. 10, pp. 1037–1041, 1997.

[15] ——, "Counterintuitive behavior of social systems," *Technological Forecasting and Social Change*, vol. 3, pp. 1–22, 1971.

[16] G. P. Richardson, "Reflections on the foundations of system dynamics," *System Dynamics Review*, vol. 27, no. 3, pp. 219–243, 2011.

[17] A. J. Lotka, "Elements of physical biology," *Science Progress in the Twentieth Century (1919-1933)*, vol. 21, no. 82, pp. 341–343, 1926.

[18] Y. Barlas, "Formal aspects of model validity and validation in system dynamics," *System Dynamics Review: The Journal of the System Dynamics Society*, vol. 12, no. 3, pp. 183–210, 1996.

[19] A. Gábor and J. R. Banga, "Robust and efficient parameter estimation in dynamic models of biological systems," *BMC systems biology*, vol. 9, no. 1, p. 74, 2015.

[20] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: new estimates of r&d costs," *Journal of health economics*, vol. 47, pp. 20–33, 2016.

[21] B. Gay and B. Dousset, "Innovation and network structural dynamics:

Study of the alliance network of a major sector of the biotechnology industry," *Research policy*, vol. 34, no. 10, pp. 1457–1475, 2005.

[22] J. Bercovitz and M. Feldman, "Entpreprenerial universities and technology transfer: A conceptual framework for understanding knowledge-based economic development," *The Journal of Technology Transfer*, vol. 31, no. 1, pp. 175–188, 2006.

[23] K. R. Fabrizio, "Absorptive capacity and the search for innovation," *Research policy*, vol. 38, no. 2, pp. 255–267, 2009.

[24] T. E. Stuart, S. Z. Ozdemir, and W. W. Ding, "Vertical alliance networks: The case of university–biotechnology–pharmaceutical alliance chains," *Research Policy*, vol. 36, no. 4, pp. 477–498, 2007.

[25] W. W. Powell, K. W. Koput, and L. Smith-Doerr, "Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology," *Administrative science quarterly*, pp. 116–145, 1996.

[26] J. Zhang, C. Baden-Fuller, and V. Mangematin, "Technological knowledge base, r&d organization structure and alliance formation: Evidence from the biopharmaceutical industry," *Research policy*, vol. 36, no. 4, pp. 515–528, 2007.

[27] C. Phelps, R. Heidl, and A. Wadhwa, "Knowledge, networks, and

knowledge networks: A review and research agenda," *Journal of Management*, vol. 38, no. 4, pp. 1115–1166, 2012.

[28] H. Singh, D. Kryscynski, X. Li, and R. Gopal, "Pipes, pools, and filters: how collaboration networks affect innovative performance," *Strategic Management Journal*, vol. 37, no. 8, pp. 1649–1666, 2016.

[29] M. A. Schilling and C. C. Phelps, "Interfirm collaboration networks: The impact of large-scale network structure on firm innovation," *Management Science*, vol. 53, no. 7, pp. 1113–1126, 2007.

[30] J. Guan and Q. Zhao, "The impact of university–industry collaboration networks on innovation in nanobiopharmaceuticals," *Technological Forecasting and Social Change*, vol. 80, no. 7, pp. 1271–1286, 2013.

[31] K. B. Whittington, J. Owen-Smith, and W. W. Powell, "Networks, propinquity, and innovation in knowledge-intensive industries," *Administrative science quarterly*, vol. 54, no. 1, pp. 90–122, 2009.

[32] M. Fritsch and M. Kauffeld-Monz, "The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks," *The Annals of Regional Science*, vol. 44, no. 1, pp. 21–38, 2010.

BIBLIOGRAPHY

[33] R. S. Burt, "Structural holes and good ideas," *American journal of sociology*, vol. 110, no. 2, pp. 349–399, 2004.

[34] R. Filieri and S. Alguezaui, "Structural social capital and innovation. is knowledge transfer the missing link?" *Journal of Knowledge Management*, vol. 18, no. 4, pp. 728–757, 2014.

[35] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[36] I. Guler and A. Nerkar, "The impact of global and local cohesion on innovation in the pharmaceutical industry," *Strategic Management Journal*, vol. 33, no. 5, pp. 535–549, 2012.

[37] A. Edelmann, J. Moody, and R. Light, "Disparate foundations of scientists' policy positions on contentious biomedical research," *Proceedings of the National Academy of Sciences*, p. 201613580, 2017.

[38] D. Lee, K. Kirkpatrick-Husk, and R. Madhavan, "Diversity in alliance portfolios and performance outcomes: A meta-analysis," *Journal of Management*, vol. 43, no. 5, pp. 1472–1497, 2017.

[39] S. K. Cohen and T. Caner, "Converting inventions into breakthrough innovations: The role of exploitation and alliance network knowledge

heterogeneity," *Journal of Engineering and Technology Management*, vol. 40, pp. 29–44, 2016.

[40] S. Rodan and C. Galunic, "More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness," *Strategic management journal*, vol. 25, no. 6, pp. 541–562, 2004.

[41] R. C. Sampson, "R&d alliances and firm performance: The impact of technological diversity and alliance organization on innovation," *Academy of management journal*, vol. 50, no. 2, pp. 364–386, 2007.

[42] A. Parkhe, "Interfirm diversity, organizational learning, and longevity in global strategic alliances," *Journal of international business studies*, vol. 22, no. 4, pp. 579–601, 1991.

[43] M. V. Tomasello, C. J. Tessone, and F. Schweitzer, "Quantifying knowledge exchange in r&d networks: a data-driven model," 2015.

[44] T. Bar and A. Leiponen, "A measure of technological distance," *Economics Letters*, vol. 116, no. 3, pp. 457–459, 2012.

[45] I. M. Cockburn and R. M. Henderson, "Scale and scope in drug development: unpacking the advantages of size in pharmaceutical research," *Journal of health economics*, vol. 20, no. 6, pp. 1033–1057, 2001.

BIBLIOGRAPHY

[46] L. Jost, "Entropy and diversity," *Oikos*, vol. 113, no. 2, pp. 363–375, 2006.

[47] D. J. Miller, "Technological diversity, related diversification, and firm performance," *Strategic Management Journal*, vol. 27, no. 7, pp. 601–619, 2006.

[48] L. Wang, A. Plump, and M. Ringel, "Racing to define pharmaceutical r&d external innovation models," *Drug discovery today*, vol. 20, no. 3, pp. 361–370, 2015.

[49] W. W. Powell, D. R. White, K. W. Koput, and J. Owen-Smith, "Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences," *American journal of sociology*, vol. 110, no. 4, pp. 1132–1205, 2005.

[50] R. Widdus, "Public-private partnerships for health: their main targets, their diversity, and their future directions," *Bulletin of the World Health Organization*, vol. 79, no. 8, pp. 713–720, 2001.

[51] M. Nieto and P. Quevedo, "Absorptive capacity, technological opportunity, knowledge spillovers, and innovative effort," *Technovation*, vol. 25, no. 10, pp. 1141–1157, 2005.

[52] V. Mangematin and L. Nesta, "What kind of knowledge can a firm

absorb?" *International Journal of Technology Management*, vol. 18, no. 3-4, pp. 149–172, 1999.

[53] M. Makri, M. A. Hitt, and P. J. Lane, "Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions," *Strategic Management Journal*, vol. 31, no. 6, pp. 602–628, 2010.

[54] L. L. Ouellette, "How many patents does it take to make a drug? follow-on pharmaceutical patents and university licensing," 2010.

[55] Desa, *World Population Prospects: The 2010 Revision, Volume II-Demographic Profiles*. UN, 2013.

[56] J. Mack, "Ipcc third assessment report, climate change 2001: Findings, criticism, and lessons for next time, the," *Colo. J. Int'l Envtl. L. & Pol'y*, vol. 17, p. 21, 2005.

[57] P. Gerland, A. E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, L. Alkema, B. K. Fosdick, J. Chunn, N. Lalic *et al.*, "World population stabilization unlikely this century," *Science*, vol. 346, no. 6206, pp. 234–237, 2014.

[58] A. E. Raftery, N. Li, H. Ševčíková, P. Gerland, and G. K. Heilig, "Bayesian

probabilistic population projections for all countries," *Proceedings of the National Academy of Sciences*, vol. 109, no. 35, pp. 13 915–13 921, 2012.

[59] T. Dietz and E. A. Rosa, "Rethinking the environmental impacts of population, affluence and technology," *Human ecology review*, vol. 1, no. 2, pp. 277–300, 1994.

[60] D. H. Meadows, D. L. Meadows, J. Randers, and W. W. Behrens, "The limits to growth," *New York*, vol. 102, 1972.

[61] H. S. Cole, *Models of doom: A critique of the limits to growth*. Universe Pub, 1973.

[62] U. Bardi, "Criticism to "the limits to growth"," in *The Limits to Growth Revisited*. Springer, 2011, pp. 49–62.

[63] M. D. Mesarovic and E. C. Pestel, "A goal-seeking and regionalized model for analysis of critical world relationships—the conceptual foundation," *Kybernetes*, vol. 1, no. 2, pp. 79–85, 1972.

[64] A. O. Herrera, H. D. Scolnik, G. Chichilnisky, G. C. Gallopin, and J. E. Hardoy, *Catastrophe or new society?: a Latin American world model*. IDRC, Ottawa, ON, CA, 1976.

[65] H. Linneman, J. d. Hoogh, M. A. Keyzer, H. D. Van Heemst *et al.*, *MOIRA: Model of International Relations in Agriculture. Report of the project*

group" Food for a doubling world population". North Holland Publishing Comp., 1979.

[66] P. Roberts, "Sarum 76—a global modelling project," *Futures*, vol. 9, no. 1, pp. 3–16, 1977.

[67] A. Onishi, "Report on project fugi: future of global interdependence," in *Fifth IIASA global modeling g conference*. IIASA Laxenburg, Austria, September 1977, 1977, pp. 1–183.

[68] W. Leontief, "Future of the world economy: a united nations study.[1980, 1990, and 2000]," 1977.

[69] W. D. Nordhaus, *Managing the global commons: the economics of climate change*. MIT press Cambridge, MA, 1994, vol. 31.

[70] S. C. Peck and T. J. Teisberg, "Ceta: a model for carbon emissions trajectory assessment," *The Energy Journal*, pp. 55–77, 1992.

[71] A. Manne, R. Mendelsohn, and R. Richels, "Merge: A model for evaluating regional and global effects of ghg reduction policies," *Energy policy*, vol. 23, no. 1, pp. 17–34, 1995.

[72] A. S. Manne, "Global 2100," 1992.

[73] H. Dowlatabadi, "Integrated assessment models of climate change: An incomplete overview," *Energy Policy*, vol. 23, no. 4-5, pp. 289–296, 1995.

[74] D. L. Kelly and C. D. Kolstad, "Integrated assessment models for climate change control," *International yearbook of environmental and resource economics*, vol. 2000, pp. 171–197, 1999.

[75] S. Motesharrei, J. Rivas, and E. Kalnay, "Human and nature dynamics (handy): Modeling inequality and use of resources in the collapse or sustainability of societies," *Ecological Economics*, vol. 101, pp. 90–102, 2014.

[76] T. S. Rowan, H. R. Maier, J. Connor, and G. C. Dandy, "An integrated dynamic modeling framework for investigating the impact of climate change and variability on irrigated agriculture," *Water Resources Research*, vol. 47, no. 7, 2011.

[77] J. Sterman, T. Fiddaman, T. Franck, A. Jones, S. McCauley, P. Rice, E. Sawin, and L. Siegel, "Climate interactive: the c-roads climate policy model," *System Dynamics Review*, vol. 28, no. 3, pp. 295–305, 2012.

[78] J. Bongaarts, "A framework for analyzing the proximate determinants of fertility," *Population and development review*, pp. 105–132, 1978.

[79] ——, "Modeling the fertility impact of the proximate determinants: Time for a tune-up," *Demographic Research*, vol. 33, p. 535, 2015.

[80] W. Lutz and K. Samir, "Global human capital: Integrating education and population," *Science*, vol. 333, no. 6042, pp. 587–592, 2011.

[81] L. J. Ralph and C. D. Brindis, "Access to reproductive healthcare for adolescents: establishing healthy behaviors at a critical juncture in the lifecourse," *Current opinion in obstetrics and gynecology*, vol. 22, no. 5, pp. 369–374, 2010.

[82] R. M. Barber, N. Fullman, R. J. Sorensen, T. Bollyky, M. McKee, E. Nolte, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas *et al.*, "Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the global burden of disease study 2015," *The Lancet*, vol. 390, no. 10091, pp. 231–266, 2017.

[83] T. Piketty, "About capital in the twenty-first century," *American Economic Review*, vol. 105, no. 5, pp. 48–53, 2015.

[84] J. Hansen, M. Sato, G. Russell, and P. Kharecha, "Climate sensitivity, sea level and atmospheric carbon dioxide," *Phil. Trans. R. Soc. A*, vol. 371, no. 2001, p. 20120294, 2013.

[85] G. Myhre, E. J. Highwood, K. P. Shine, and F. Stordal, "New estimates of radiative forcing due to well mixed greenhouse gases," *Geophysical research letters*, vol. 25, no. 14, pp. 2715–2718, 1998.

[86] M. A. Hanjra and M. E. Qureshi, "Global water crisis and future food security in an era of climate change," *Food Policy*, vol. 35, no. 5, pp. 365–377, 2010.

[87] M. L. Parry, C. Rosenzweig, A. Iglesias, M. Livermore, and G. Fischer, "Effects of climate change on global food production under sres emissions and socio-economic scenarios," *Global Environmental Change*, vol. 14, no. 1, pp. 53–67, 2004.

[88] S. Solomon, G.-K. Plattner, R. Knutti, and P. Friedlingstein, "Irreversible climate change due to carbon dioxide emissions," *Proceedings of the national academy of sciences*, pp. pnas–0 812 721 106, 2009.

[89] N. W. Arnell, "Climate change and global water resources," *Global environmental change*, vol. 9, pp. S31–S49, 1999.

[90] H. S. Houthakker, "An international comparison of household expenditure patterns, commemorating the centenary of engel's law," *Econometrica, Journal of the Econometric Society*, pp. 532–551, 1957.

[91] I. A. Shiklomanov, "Appraisal and assessment of world water resources," *Water international*, vol. 25, no. 1, pp. 11–32, 2000.

[92] L. Scrucca, "On some extensions to ga package: hybrid optimisation,

parallelisation and islands evolution," *arXiv preprint arXiv:1605.01931*, 2016.

[93] D. J. Bogue, G. Liegel, M. Kozloski *et al.*, "Immigration, internal migration, and local mobility in the us," *Books*, 2009.

[94] R. Black, S. R. Bennett, S. M. Thomas, and J. R. Beddington, "Climate change: Migration as adaptation," *Nature*, vol. 478, no. 7370, pp. 447–449, 2011.

[95] M. E. Hauer, J. M. Evans, and D. R. Mishra, "Millions projected to be at risk from sea-level rise in the continental united states," *Nature Climate Change*, vol. 6, no. 7, pp. 691–695, 2016.

[96] M. E. Hauer, "Migration induced by sea-level rise could reshape the us population landscape," *Nature Climate Change*, vol. 7, pp. 321–325, 2017.

[97] D. Dodman, "Blaming cities for climate change? an analysis of urban greenhouse gas emissions inventories," *Environment and urbanization*, vol. 21, no. 1, pp. 185–201, 2009.

[98] G. Marland, T. A. Boden, R. J. Andres, A. Brenkert, and C. Johnston, "Global, regional, and national fossil fuel co2 emissions," *Trends: A compendium of data on global change*, pp. 34–43, 2003.

[99] N. Y. Chan, K. L. Ebi, F. Smith, T. F. Wilson, and A. E. Smith, "An integrated assessment framework for climate change and infectious diseases." *Environmental Health Perspectives*, vol. 107, no. 5, p. 329, 1999.

[100] Y. Barlas, "System dynamics: systemic feedback modeling for policy analysis," *System*, vol. 1, no. 59, 2007.

[101] P. Backlund, A. Janetos, D. Schimel *et al.*, "The effects of climate change on agriculture, land resources, water resources, and biodiversity in the united states." *The effects of climate change on agriculture, land resources, water resources, and biodiversity in the United States.*, 2008.

[102] U. Strasser, U. Vilsmaier, F. Prettenhaler, T. Marke, R. Steiger, A. Damm, F. Hanzer, R. A. Wilcke, and J. Stötter, "Coupled component modelling for inter-and transdisciplinary climate change impact research: Dimensions of integration and examples of interface design," *Environmental modelling & software*, vol. 60, pp. 180–187, 2014.

[103] S. L. Cutter, L. Barnes, M. Berry, C. Burton, E. Evans, E. Tate, and J. Webb, "A place-based model for understanding community resilience to natural disasters," *Global environmental change*, vol. 18, no. 4, pp. 598–606, 2008.

[104] J. M. Links, B. S. Schwartz, S. Lin, N. Kanarek, J. Mitrani-Reiser, T. K.

Sell, C. R. Watson, D. Ward, C. Slemp, R. Burhans *et al.*, "Copewell: a conceptual framework and system dynamics model for predicting community functioning and resilience after disasters," *Disaster medicine and public health preparedness*, vol. 12, no. 1, pp. 127–137, 2018.

[105] G. Luber and M. McGeehin, "Climate change and extreme heat events," *American journal of preventive medicine*, vol. 35, no. 5, pp. 429–435, 2008.

[106] G. A. Meehl and C. Tebaldi, "More intense, more frequent, and longer lasting heat waves in the 21st century," *Science*, vol. 305, no. 5686, pp. 994–997, 2004.

[107] U. Franck, M. Krüger, N. Schwarz, K. Grossmann, S. Röder, and U. Schlink, "Heat stress in urban areas: Indoor and outdoor temperatures in different urban structure types and subjectively reported well-being during a heat wave in the city of leipzig," *Meteorologische Zeitschrift*, vol. 22, no. 2, pp. 167–177, 2013.

[108] J. L. Nguyen, J. Schwartz, and D. W. Dockery, "The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity," *Indoor air*, vol. 24, no. 1, pp. 103–112, 2014.

[109] M. McCormack, A. Scott, B. Zaitchik, Z. He, and R. Peng, "Indoor

heat exposure: Links to weather and housing characteristics." Presented as the Seventh American Meteorological Society Conference on Environment and Health in New Orleans, Louisiana, 2016.

[110] A. Tasneem, L. Aberle, H. Ananth, S. Chakraborty, K. Chiswell, B. J. McCourt, and R. Pietrobon, "The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty," *PloS one*, vol. 7, no. 3, p. e33677, 2012.

[111] M. Smithson and J. Verkuilen, "A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables." *Psychological methods*, vol. 11, no. 1, p. 54, 2006.

Vita

Gary Lin was born in Los Angeles, California, USA on September 16, 1989. He was raised in Clarendon Hills, Illinois, USA by his parents, Tsai Hua (Teresa) Lee and Yuan Chi (Joseph) Lin. He attended Hinsdale Central Highschool and ran Track and Cross Country where he earned "Red Devil Award" for outstanding character and service. Later, he attended the University of Colorado at Boulder where he earned a Bachelor of Science in Civil Engineering and Bachelor of Arts in Economics. He was also the Co-captain of Steel Bridge Design Team, Vice President of the Student Chapter of the American Society of Civil Engineers (ASCE), and earned First Place for his Senior Design Capstone Project. After college, he spent a brief time working as a logistics coordinator at Ferrara Candy Company.

His intellectual passion was always in the intersection of engineering and social sciences. When he began his doctoral studies at the Johns Hopkins University in 2013, Gary was introduced to the systems approach and modeling. Gary attempted many ambitious modeling projects as well as unconventional

projects with highly interdisciplinary teams of engineers, law experts, social scientists, public health experts, engineers, and designers. His work involved many collaborators like the MIT Collaborative Initiatives, Bill and Melinda Gates Institute for Population and Reproductive Health, and the Bloomberg American Health Initiative. His other interest is in the field of sustainability led him to earn a Masters of Science in Engineering in Geography and Environmental Engineering at Johns Hopkins. Recently, he is passionate about spreading systems methodology in the field of public health and medicine. In 2018, Gary earned his Doctor of Philosophy in Civil Engineering with a focus in systems science. Gary will continue his research and learning as a Postdoctoral Research Associate at the Department of Emergency Medicine at Johns Hopkins School of Medicine. Gary also enjoys playing a round of disc golf with his friends.